# UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE PSICOLOGÍA

Departamento de Metodología de las Ciencias del Comportamiento


## TESIS DOCTORAL

Equivalencia e invarianza de medida entre grupos: análisis factorial confirmatorio vs teoría de respuesta al item

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Patricia Recio Saboya

Directores

Carmen Santisteban Requena
Jesús Mª Alvarado Izquierdo

**Madrid, 2012**

Departamento de Metodología de las ciencias del Comportamiento
Facultad de Psicología
Universidad Complutense de Madrid

Programa de Doctorado en Psicología *experimental y aplicada: Atención, aprendizaje y percepción*


# *Equivalencia e invarianza de medida entre grupos: análisis factorial confirmatorio vs teoría de respuesta al ítem*

Doctorando: Patricia Recio Saboya

Directores de tesis: Carmen Santisteban Requena y Jesús Mª Alvarado Izquierdo

Madrid, 2012

*A Juan Miguel, Marco y Tania.*

# Agradecimientos

La única ventaja de alargar esta tarea más de lo debido es que, al final, son más las personas que acaban participando en ella. Muchas son las que me han ayudado a sacar adelante este trabajo, más de las que podría citar aquí.

Debo agradecer a mis directores de tesis, Carmen Santisteban y Jesús Alvarado quienes, a pesar de otras muchas ocupaciones y dificultades, se comprometieron para sacar esta tesis adelante. Gracias por vuestra dedicación y paciencia. Los informes realizados por los profesores Sergio Escorial y Miguel Ángel Mateo han contribuido también a terminar de pulir el manuscrito final. Del resto de errores y erratas que todavía queden en el trabajo, me temo que seré la única responsable.

Siempre he pensado que todos necesitamos un referente para hacer las cosas mejor. El mío ha sido, sin duda, Mª José Navas. He aprendido muchísimo trabajando con ella, pero lo que más valoro y admiro es el tesón, la energía y la honestidad con la que afronta cada proyecto. Gracias por todo.

A mis compañeros del Dpto. de Metodología de las Ciencias del Comportamiento de la UNED, con especial mención a Laura Quintanilla, José Mª Merino, José Manuel Reales, Encarnación Sarriá, Juan Carlos Suárez y Pablo Holgado, por aconsejarme e infundirme ánimos.

No puedo olvidarme de citar a algunos compañeros y amigos de dentro y fuera de la Facultad de Psicología de la UNED, como Antonio Contreras, Begoña Delgado,

Isabel Gómez, Pilar del Pozo, Inmaculada Sánchez, Chema Luzón, Fernando Molero, Cristina García, Esther Ramos y Eva Mª de la Peña.

Quiero agradecer a mis padres, Antonio y Nieves, por apoyarme en todas las decisiones que he tomado a lo largo de mi vida. A mi familia política, en especial a Juanjo y Petri, por su cariño. A mis hermanos, Marcos, Óscar, Rosa, Alejandro y Adrián, por todos los buenos momentos que pasamos juntos. Sin duda, ellos me enseñaron a reírme de mi misma.

Gracias a mis hijos, Marco y Tania, por cambiar mis esquemas mentales, enseñándome la importancia real de las cosas. Gracias a Juan Miguel, mi amor y compañero, por tanta felicidad compartida. Hemos caminado juntos de la primera a la última página de este trabajo, y el apoyo y confianza que siempre ha depositado en mí han resultado claves en los momentos difíciles. Más que nadie ha esperado este momento.

Madrid, Marzo de 2012

# Índice

## PRESENTACIÓN

Cuando un investigador administra un test psicológico a individuos que pertenecen a diferentes grupos asume, de alguna manera, que ese test está midiendo el mismo constructo bajo diferentes condiciones (McDonald, 1999). Las poblaciones de interés habitualmente se definen por variables demográficas, como el sexo, la edad, la raza, el país de origen o el idioma. Estas condiciones incluyen la estabilidad de la medida en diferentes culturas (Riordan y Vandenberg, 1994), evaluadores (Facteau y Craig, 2001), o en distintos procedimientos de administración de la prueba (Taris, Bok y Meijer, 1998).

Si las puntuaciones del test se utilizan para comparar directamente a estos grupos, para que estas comparaciones tengan validez es necesario que el test mida el mismo constructo en cada grupo y que la relación entre las puntuaciones del test y las puntuaciones en el constructo sean invariantes o equivalentes en las distintas poblaciones. Por lo tanto, la cuestión de la equivalencia de medida es relevante para prácticamente la totalidad del empleo de las puntuaciones de tests en distintas poblaciones, cuando el objetivo es simplemente evaluar diferencias individuales o utilizar las puntuaciones del test como predictores (Millsap, 1997, 2011).

Se dice que un test posee *equivalencia y/o invarianza de medida* entre grupos cuando individuos con idéntico nivel de rasgo en el constructo medido, pero que pertenecen a distintos grupos, tienen la misma probabilidad de obtener igual puntuación en el test (Meredith y Millsap, 1992). Por el contrario, violaciones en la equivalencia de

medida implican que dos individuos con idéntico nivel en el constructo y que proceden de distintos grupos, tengan una puntuación esperada diferente en el test.

Existen en la actualidad, una gran variedad de métodos estadísticos para investigar si un test posee la propiedad de equivalencia de medida (ver por ejemplo, Steenkamp y Baumgartner, 1998; Vandenberg y Lance, 2000; Widaman y Reise, 1997). Todos estos métodos requieren que, de una u otra manera, los individuos de diferentes poblaciones se emparejen en el constructo de interés, para realizar comparaciones de las puntuaciones observadas dentro de los grupos equiparados (Millsap y Kwok, 2004).

Una estrategia general para realizar la equiparación es ajustar un modelo de medida a los datos de cada población y después evaluar si la forma de las relaciones entre la variable latente y las puntuaciones observadas es la misma en las distintas poblaciones. En este trabajo se utilizan modelos basados en dos aproximaciones diferentes para evaluar la equivalencia de medida: Análisis Factorial Confirmatorio (AFC) y Teoría de Respuesta al Ítem (TRI).

Los modelos de AFC son centrales en los modelos de ecuaciones estructurales. Su desarrollo comenzó en la década de los 70, principalmente a cargo de sociólogos y economistas (ver, por ejemplo, Jöreskog, 1971, 1973; McArdle y McDonald, 1984). El modelado de ecuaciones estructurales expande el análisis factorial exploratorio tradicional a uno confirmatorio y permite combinarlo con un componente estructural, especificando las relaciones entre los elementos que lo configuran.

Para valorar la equivalencia de la medida entre grupos se utiliza una condición en la que se fuerza a los parámetros del modelo factorial a tener los mismos valores en las distintas poblaciones. Un modelo factorial típico incluye muchos parámetros, por lo que se pueden establecer muchas fuentes potenciales de violación de equivalencia. El AFC proporciona tests estadísticos para subconjuntos de estos parámetros, o para todos los parámetros simultáneamente. Se han propuesto secuencias organizadas de estas pruebas de equivalencia en etapas sucesivas, cada una de las cuales englobaría las anteriores, en una serie de modelos anidados cada vez más restrictivos (ver por ejemplo, Jöreskog, 1971; Steenkamp y Baumgartner, 1998; Vandenberg y Lance, 2000; Widaman y Reise, 1997).

Los modelos de la TRI definen relaciones probabilísticas y no lineales entre los constructos hipotéticos y sus indicadores. Se desarrollaron en Psicología y Educación desde la década de los 60 (ver Lord y Novick, 1968). En su comienzo se diseñaron casi exclusivamente para la evaluación de las aptitudes, pero a partir de la década de los 90 su utilización en la evaluación de actitudes se ha incrementado (Embretson y Reise, 2000).

Desde la TRI se utilizan diferentes estrategias para valorar la equivalencia de medida entre grupos, habitualmente enfocadas al funcionamiento diferencial de los ítems (Differential Item Functioning, DIF). En esta investigación se utiliza la comparación de modelos mediante el estadístico de razón de verosimilitud (Thissen, Steinberg y Gerrard, 1986; Thissen, Steinberg y Wainer, 1988, 1993) porque, al igual que en el AFC, evalúa el modelo que mejor se ajusta a los datos, en este caso comparando el modelo de ausencia de DIF con otros donde se asume que en uno o más

ítems el DIF está presente. Recurrimos además al procedimiento DFIT (Raju, van der Linden y Fleer, 1995), basado en el concepto de puntuación verdadera del test, porque está desarrollado específicamente para evaluar el funcionamiento diferencial no solo a nivel de ítem, sino además a nivel de test.

Esta memoria de tesis aborda la equivalencia de medida desde la óptica del test completo o subtests porque las decisiones relacionadas con variables psicológicas se basan frecuentemente en puntuaciones obtenidas en conjuntos de ítems y no en ítems particulares. Por ejemplo, en selección de personal, las decisiones para un puesto determinado se apoyan sobre aptitudes o actitudes que se evalúan mediante tests. De forma similar, la relación entre dos variables (por ejemplo satisfacción en el trabajo y rendimiento en el puesto) se examinan evaluando la asociación de dos tests o conjuntos de ítems. Por tanto, las propiedades de medida de las puntuaciones de los tests son de una importancia fundamental debido a las decisiones e inferencias que se basan en esas puntuaciones (Drasgow, 1995a, Navas, 2001). Sin embargo, las propiedades métricas de ítems individuales no son de una importancia directa en los casos en los que las decisiones se basan en las puntuaciones totales del test; por este motivo, en este estudio se consideran las características de los ítems particulares importantes o no en función de su contribución a las propiedades de medida del test o escala en su conjunto.

En esta misma línea de razonamiento, Drasgow (1987) y Drasgow y Hulin (1990) argumentan que el sesgo de medida debería examinarse a nivel del test completo. Por ejemplo, Drasgow (1987) encuentra que una alta proporción de ítems del *American College Testing* (ACT) presentan funcionamiento diferencial significativo entre hombres y mujeres. Sin embargo, cuando se comparan las curvas características del test

-el número esperado de puntuaciones correctas computado como una función del rasgo latente evaluado por el test- solo se encuentran diferencias triviales.

Para estudiar los diferentes procedimientos propuestos para comprobar la equivalencia métrica se utilizó un instrumento de medida de la impulsividad, el test BIS, test que fue adaptado para servir tanto a población preadolescente como adolescente (Recio, Santisteban y Alvarado, 2004). La equivalencia métrica de este test se contrastó entre muestras que diferían en género y edad, utilizando los procedimientos de AFC, test de razón de verosimilitud y el procedimiento DFIT.

Esta memoria de tesis se encuadra dentro de una investigación más amplia dirigida por la Dra. Carmen Santisteban sobre el estudio de la agresividad y otros conceptos relacionados como la impulsividad en niños y adolescentes, en la que se estableció como un objetivo clave para poder realizar las comparaciones entre las distintas subpoblaciones el estudio de la equivalencia métrica. En este sentido, el equipo de investigación en el que me integro ha publicado los resultados del análisis de la equivalencia métrica de una medida de agresividad (Santisteban, Alvarado y Recio, 2007) en el que se utilizan procedimientos de validez de constructo basados en el AFC. En la presente investigación, continuando con este esfuerzo se analizarán en profundidad los principales procedimientos para el análisis de la equivalencia métrica, discutiendo las ventajas e inconvenientes de las distintas alternativas.

# Sección I. MARCO TEÓRICO

# 1. MEDICIÓN EN PSICOLOGÍA

La medición en psicología ha sido tema de debate desde sus albores, y debe su importancia a que una psicología científica depende en buena parte de una medida (Cattell, 1981; Kline, 1998). Esto, que sucede en todas las disciplinas científicas porque, parafraseando a Cattell (1893) "la historia de la ciencia es la historia de la medida" (citado en Santisteban y Alvarado, 2001), tiene una dificultad añadida en el caso de la psicología: su objeto de estudio.

La definición de medición más utilizada en Psicología es la proporcionada por Stevens (1951), que la define como "la asignación de números a objetos o eventos de acuerdo a una regla". Esta definición, sin embargo, no está exenta de críticas desde su origen, ya que la Psicología surgió en un ambiente de corte claramente positivista, en el que se intentaba utilizar para la Psicología el mismo marco de las magnitudes físicas, esto es, dentro de lo que Savage y Ehrlich (1990) denominan como *concepción conservadora de la medición*, formalizada por los axiomas de cantidad de Hölder (una explicación detallada de los problemas en los inicios de la medición en psicología y las soluciones adoptadas puede encontrarse en Muñiz, 2001 y Navas, 1997).

La medición de constructos psicológicos es fundamental, tanto para la investigación psicológica como para la práctica profesional. Los constructos psicológicos suelen ser conceptualizados como variables latentes que subyacen al comportamiento. Como han señalado Cronbach y Meehl (1955) los constructos psicológicos son construcciones teóricas para explicar la consistencia del comportamiento en diversos contextos. Esta

conceptualización de un constructo psicológico tiene varias implicaciones para la medición: la posición de una persona sobre un constructo psicológico determinado debe ser inferido a partir de su comportamiento. Por razones prácticas, la medición de las personas en un contexto natural en psicología es la excepción y no la regla por lo que, en su lugar, se han desarrollado tests para observar respuestas. Una medición adecuada implica repetición de situaciones o ítems, por lo que las mediciones psicológicas habitualmente constan de múltiples ítems o tareas que varían en contenido. Dado que los constructos se miden a través de ítems, se debe observar una consistencia de la conducta a través de esos ítems (Embretson, 2006). Habitualmente se utilizan dos perspectivas diferentes para obtener mediciones de constructos a partir de las respuestas a una prueba: la Teoría Clásica de los Tests (TCT) y la Teoría de Respuesta a los Ítems (TRI).

La psicología como ciencia, por tanto, se ha caracterizado por el desarrollo de una extensa colección de mediciones y tests. La interpretación de las puntuaciones de los tests y las subsiguientes decisiones basadas en esas interpretaciones requieren inferencias desde las puntuaciones observadas del test hasta el constructo inobservable representado por los ítems del test (Crocker, 2006).

El consumo de tests en nuestra sociedad es grande y las consecuencias para las personas evaluadas también son a menudo importantes: acceder a la enseñanza universitaria, aprobar una oposición, tener carnet de conducir, acceder a un puesto de trabajo o a un ascenso son algunos de los ejemplos que pueden afectar directamente a la vida de las personas. Esto propició que, a partir de los años 60, hubiera una gran preocupación, no solo entre los especialistas sino en el público en general, por la posibilidad de que algunos tests psicológicos pudieran estar sesgados, favoreciendo -y por

ende perjudicando- a un grupo particular de examinados. El germen de esta preocupación se originó en EEUU y eclosionó con un estudio de Jensen (1969) que considera que la inteligencia es hereditaria y que, las diferencias que se observaban entre grupos raciales eran atribuibles a la genética. Esta afirmación de carácter genético chocó frontalmente con la opinión de los ambientalistas que defendían que la explicación de las diferencias entre grupos hay que buscarla en el posible sesgo cultural de los tests de inteligencia.

Desde ese momento, la cuestión de cómo se puede demostrar que una escala mide el mismo constructo, de la misma manera, cuando se administra a dos o más grupos distintos, ha motivado un número creciente de investigaciones en los últimos años (Cheung y Rensvold, 1999). La pregunta a la que se quiere dar respuesta desde distintos ámbitos es la siguiente: ¿son las puntuaciones de las personas que pertenecen a diferentes grupos o poblaciones comparables en la misma escala de medida? (Reise, Widaman y Pugh, 1993).

Para comparar grupos de individuos en cuanto a sus niveles en algún constructo, o respecto a las relaciones entre esos constructos, se debe asumir que los instrumentos utilizados en la evaluación tienen "equivalencia de medida" o "invarianza" entre los grupos (Drasgow, 1987). De no ser así, las diferencias entre los grupos en medias o en los patrones de las correlaciones son potencialmente artificiales y pueden ser sustantivamente erróneas.

Por lo tanto, demostrar la equivalencia de medida en diferentes grupos es crucial para avanzar en muchos ámbitos.

# 2. EQUIVALENCIA DE MEDIDA Y CONCEPTOS RELACIONADOS

## 2.1. CONCEPTO DE EQUIVALENCIA O INVARIANZA DE MEDIDA

La definición más utilizada del término equivalencia es la proporcionada por Drasgow y Kanfer (1985), según la cual un test o una subescala posee invarianza, o equivalencia de medida en varios grupos o poblaciones si personas con puntuaciones idénticas en el rasgo latente subyacente tienen la misma puntuación esperada a nivel de ítem, a nivel de puntuación total en la escala o ambos.

Una definición más formal del término se enuncia de la siguiente manera: Supongamos un conjunto de *n* mediciones *y*, obtenidas de una muestra aleatoria de sujetos. Supongamos además que estas mediciones son una función estadística de otro conjunto de *p* variables aleatorias $\theta$. Considerando, además, una variable *x* que indica el grupo (o población) al que pertenece el sujeto, podremos afirmar que nuestro conjunto de mediciones *y* es invariante o equivalente con respecto a *x* si:

$$\text{Prob}\left(y|\theta = t,\ X = x\right) = \text{Prob}\left(y|\theta = t\right) \qquad (1)$$

para todos los valores de *x* y *t*. Esto es, si la probabilidad de observar un conjunto de mediciones *y* (un conjunto de variables dependientes) para un nivel fijo de predictores $\theta = t$, es independiente del grupo al que pertenezca el sujeto. En otras palabras, un conjunto de mediciones *y* es invariante con respecto a *x* si la relación entre *y* y $\theta$, dada por Prob ($y \mid \theta = t$) es la misma con independencia del grupo al que pertenezca el sujeto. Esta definición goza también de amplio consenso (Maydeu-Olivares, Morera y Zurilla, 1998; Meredith, 1993; Millsap y Everson, 1993) y, aunque está expresada de manera formal, sigue siendo muy general: las mediciones (variables dependientes) *y* y las variables independientes $\theta$

pueden ser unidimensionales o multidimensionales, así como continuas o categóricas, y su relación dada por Prob ($y \mid \theta = t$) puede ser lineal o no lineal.

A partir de ambas definiciones es fácil concluir que, en caso de falta de equivalencia de la medida en los grupos, es equívoco compararlas. Esto es, las diferencias encontradas pueden reflejar tanto diferencias verdaderas entre los grupos, como una diferencia en la relación entre la variable latente y la puntuación observada que no es igual en ambos grupos.

En este sentido, la cuestión central de la invarianza/equivalencia de la medida radica en comprobar que bajo diferentes condiciones de observación y estudio del fenómeno, el instrumento de medida realmente mide el mismo constructo. Si no hay evidencia de presencia o ausencia de invarianza de medida (que es lo más usual) o hay evidencia de que tal invarianza no se obtendrá, entonces las bases científicas para la inferencia serán muy escasas: los hallazgos de diferencias entre individuos y grupos no podrán ser interpretados de forma inequívoca (Horn y McArdle, 1992; Millsap, 2011).

## 2.2. CONCEPTOS RELACIONADOS

En la literatura se utilizan algunos términos, que si bien no corresponden exactamente al mismo concepto de equivalencia de medida, sí tienen una clara relación con el mismo. Entre ellos cabe destacar sesgo y Funcionamiento Diferencial del Ítem (en adelante DIF, según acrónimo inglés), conceptos íntimamente ligados. Los primeros estudios de DIF empezaron en la década de los 60 bajo la denominación de sesgo. En la década de los 80 se cambió la terminología de sesgo a DIF, por motivos que eran más

políticos o linguísticos que psicométricos, ya que la razón fundamental fue que la palabra sesgo conlleva connotaciones negativas, siendo sinónima en los diccionarios de términos como perjuicio y parcialidad (Raju y Ellis, 2000).

### 2.2.1. SESGO

Habitualmente se cita como primera investigación sobre el sesgo de los ítems el trabajo de Eells, Davis, Havighurst, Herrick y Tyler (1951) (Fidalgo, 1996; McIntire y Miller, 2007). En la década de los 60, los especialistas en medida, los investigadores y el público general se han interesado y preocupado de manera creciente con la posibilidad de que la medida psicológica "trabaje de forma diferente" o esté sesgada a favor o en contra de un grupo particular de examinados. Esta creciente preocupación surgió con el movimiento de los derechos civiles en EEUU, ya que en muchas de las situaciones en las que se reivindica igualdad de derechos y oportunidades, como admisión de alumnos en educación superior y selección de personal, se utilizaban tests para tomar este tipo de decisiones.

El artículo de Jensen (1969) *How much can we boost IQ and scholastic achievement?* contribuyó a pasar de la preocupación a la polémica al considerar el componente genético de la inteligencia para justificar las diferencias raciales. Así, los genetistas defendían que las diferencias encontradas en los tests reflejaban diferencias reales en las aptitudes, mientras que los ambientalistas defendían que estas diferencias se debían a que los test estaban sesgados en contra de los grupos minoritarios.

De este modo, la mayoría de las preocupaciones sobre el sesgo de los tests se centran históricamente en el rendimiento diferencial en función del sexo o la raza. Si las puntuaciones medias en el test de estos grupos (los hombres frente a las mujeres o los negros frente a los blancos) son diferentes, entonces se plantea la cuestión de si esta diferencia refleja o no sesgos de la prueba aplicada. En este contexto, los primeros métodos para evaluar el sesgo de los ítems (1) focalizan su atención en comparaciones de solo dos grupos de sujetos, (2) utilizan la terminología de grupo focal y grupo de referencia para denotar al grupo minoritario y mayoritario respectivamente y (3) analizan ítems dicotómicos casi exclusivamente.

La literatura estadística en tests psicológicos distingue entre, al menos, dos formas posibles de sesgo entre grupos: el sesgo predictivo o externo y el sesgo de medida o interno (Camilli y Shepard, 1994; Cole, 1981; Drasgow, 1982, 1987; Jensen, 1980; Reynolds y Brown, 1984). El sesgo externo sucede cuando existen diferencias de grupo en la relación entre el test y un criterio externo, o lo que es lo mismo, las puntuaciones del test tienen diferentes correlaciones con variables externas al test para dos o más grupos de examinados. Por tanto, en tests sesgados, las ecuaciones de regresión del criterio externo sobre las puntuaciones en el test calculadas en diferentes grupos serán diferentes. Esto ocasiona una validez predictiva diferencial de la medida. Esta validez diferencial o pérdida de invarianza puede ser un motivo de preocupación dependiendo del contexto en el que se utilice el test. En el contexto de muchas investigaciones la predicción diferencial de una medida se anticipa por una teoría sustantiva y puede ser una cuestión central en una investigación.

Una segunda forma de sesgo ocurre cuando las relaciones internas de un test (por ejemplo, las covarianzas entre las respuestas a los ítems) difieren en los dos o más grupos de examinados. Los procedimientos para evaluar el sesgo interno utilizan la puntuación total en el test como criterio para juzgar las diferencias entre grupos. En palabras de Millsap (1995, pag. 577), "el sesgo de medida se refiere a las diferencias de grupo en la relación entre el test y la variable latente que se mide" Esta denominación de *sesgo de medida* se relaciona, de forma inversa, con la definición de equivalencia que se maneja en este trabajo, ya que, de existir, el sesgo de medida tiene como consecuencia que no haya invarianza o equivalencia de medida en los grupos.

El ambiente altamente politizado ha contribuido a la controversia de la que el término sesgo ha sido objeto en la literatura (Jensen, 1980) debida, en gran parte, a la utilización de esta misma palabra con dos significados: por una parte el significado y las connotaciones sociales de la palabra sesgo y por otra su significado estadístico.

Por este motivo, a raíz de la publicación de Holland y Thayer (1988) para la acepción estadística se ha ido sustituyendo el término sesgo por otro más preciso: funcionamiento diferencial de los ítems.

La palabra sesgo se reserva ahora a las situaciones en que se puede establecer relación entre el funcionamiento diferencial y el constructo que se pretende medir (Camilli, 1993; Shealy y Stout, 1993). Por lo tanto, sólo se puede hablar de sesgo en términos de validez de constructo. Decir que un ítem/test está sesgado implica necesariamente un funcionamiento diferencial entre grupos, pero indica además que no mide lo que pretende

medir o que mide más cosas de las que pretende medir (Fidalgo, 1996; Gómez, Hidalgo y Guilera, 2010).

### 2.2.2. EQUIDAD

El concepto de equidad surgió en la consideración de las diferencias de género y raza en el salario (Millsap y Meredith, 1994) y ha sido fundamental en la administración de las pruebas educativas y psicológicas en el último medio siglo. Dentro del ámbito de aplicación de la prueba, el término equidad puede asumir una serie de significados relacionados con la forma en que las puntuaciones del test o de los ítems se utilizan para evaluar a los sujetos en decisiones de selección y clasificación. De acuerdo con los últimos Estándares de los Tests Psicológicos y Educativos (AERA, APA y NCME, 1999) la equidad en la prueba puede ser interpretada en relación con la falta de sesgo a nivel de ítem o del test, en relación con un tratamiento equitativo en el proceso de evaluación y en relación a la igualdad de oportunidades para aprender de todos los grupos de evaluados (Penfield y Camilli, 2007).

### 2.2.3. FUNCIONAMIENTO DIFERENCIAL DEL ÍTEM (DIF)

Este concepto tiene su origen en la TRI, de ahí la definición de algunos autores (Hambleton y Swaminathan, 1985; Hambleton, Swaminathan y Rogers, 1991; Lord, 1980) que consideran que un ítem presenta DIF cuando sujetos de distintos grupos, que tienen el mismo nivel en el rasgo o atributo evaluado por el ítem, tienen diferentes probabilidades de responder correctamente al ítem o tienen diferentes puntuaciones esperadas en el ítem.

Desde finales del siglo pasado se ha puesto de manifiesto la importancia de analizar las causas del DIF (Gómez, Hidalgo y Guilera, 2010). En este sentido, Ferne y Rupp (2007) en una revisión de 27 estudios que intentan identificar las causas del DIF constatan que los avances logrados son poco relevantes. Una perspectiva que puede resultar fructífera a la hora de analizar los motivos por los que sucede el funcionamiento diferencial es la perspectiva multidimensional, que considera que el DIF se produce cuando hay ítems multidimensionales en un test que pretende ser unidimensional y existen diferentes distribuciones entre grupos en alguno de los constructos que no se pretenden medir (Ackerman, 1992; Shealy y Stout, 1993).

En esta línea, Haladyna y Downing (2004), retomando las ideas de Messick (1989), denominan varianza irrelevante del constructo a todos los factores de personalidad y situacionales que influyen en la puntuación que se obtiene en un ítem o test pero no forman parte del rasgo que se desea medir. Consideran, además, que la varianza irrelevante del constructo es una gran amenaza a la validez de las puntuaciones de los test porque constituye un error sistemático. Su perspectiva es integradora, considerando una taxonomía para el estudio de los errores asociados con la varianza irrelevante de constructo que incluye 21 fuentes de error organizadas en las distintas fases de trabajo con un test (preparación, desarrollo del test, administración y puntuación).

En la literatura se distingue entre DIF uniforme y no uniforme (Mellenbergh, 1982). Se habla de DIF uniforme cuando la probabilidad de contestar correctamente al ítem es mayor para un grupo que para otro consistentemente a lo largo de todos los niveles del rasgo, es decir, cuando no existe interacción entre el nivel de rasgo y el grupo de pertenencia. En cambio, en el DIF no uniforme la diferencia en las probabilidades de

responder correctamente al ítem en los dos grupos no es la misma en todos los niveles del rasgo, hay, por tanto, una interacción entre el nivel de rasgo y la pertenencia a un determinado grupo.

Existen numerosos estudios que comparan los diversos métodos para detectar DIF (ver, por ejemplo, Camilli y Shepard, 1994; Fidalgo, 1996; Gómez e Hidalgo, 1997; Hidalgo y Gómez, 1999; Hidalgo y López, 2000; Holland y Wainer, 1993; Millsap y Everson, 1993; Penfield y Camilli, 2007; Potenza y Dorans, 1995; Thissen *et al*., 1993).

Estos estudios han dado lugar a múltiples clasificaciones de los procedimientos para detectar DIF. En líneas generales, las diferencias entre los distintos procedimientos se basan en los siguientes criterios: (a) son parámetricos o no paramétricos; (b) se basan en variables latentes u observadas; (c) modelizan rasgos unidimensionales y/o multidimensionales; (d) detectan DIF uniforme y/o no uniforme; (e) examinan respuestas dicotómicas y/o politómicas; (f) incluyen covariables en el modelo y (g) utilizan o no una variable grupo.

En términos generales estos procedimientos se pueden dividir en dos amplias categorías: 1) los que utilizan como criterio de equiparación de los grupos la puntuación observada en el test -estadístico Mantel-Haenszel (Holland and Thayer, 1988), estandarización (Dorans y Kullick, 1986), modelos de regresión logística (Swaminathan y Rogers, 1990), modelos loglineales (Mellenbergh, 1982), análisis discriminante logístico (Miller y Spray, 1993), método delta-plot (Angoff y Ford, 1973)- y los que utilizan el rasgo latente estimado bajo algún modelo de TRI -estadístico de Lord (Lord, 1980), medidas de área (Raju, 1988, 1990), métodos basados en la comparación de modelos (Thissen, *et al.*, 1988; 1993), SIBTEST(Shealy y Stout, 1993)-

Probar la efectividad de estas técnicas bajo distintas manipulaciones de DIF (porcentaje de ítems con DIF en el test, cantidad de DIF, tipo de DIF, presencia o no de impacto entre grupos, tamaño muestral de los grupos bajo estudio, distintos formatos de respuesta de los ítems, presencia de multidimensionalidad) ha sido y es una de las tareas actuales de los psicómetras, con la finalidad de proporcionar al profesional interesado información relevante para seleccionar un procedimiento de detección de DIF (un resumen de las ventajas e inconvenientes de estas técnicas aparece en Gómez e Hidalgo, 1997 y en Hidalgo y Gómez, 1999).

La mayoría de los métodos más extendidos se han aplicado a la detección de DIF en ítems de respuesta dicotómica, es decir, aquellos ítems que poseen dos categorías de respuesta y el sujeto debe elegir una de ellas. Sin embargo, en la medición psicológica, una gran parte de los tests requieren un formato de respuesta con más de dos categorías, tal y como es el caso de la parte empírica de esta investigación.

La forma de proceder en la evaluación del DIF en ítems politómicos es paralela a la forma de proceder en el caso dicotómico: se trata de valorar si la probabilidad de elegir una determinada categoría de un ítem para sujetos con idéntico nivel en la característica evaluada varía o no según el grupo al que pertenece el sujeto.

Muchas de las técnicas propuestas para ítems politómicos son generalizaciones del caso dicotómico, existiendo otras que sirven para cualquier tipo de ítem. Así, se utilizan métodos basados en la TRI (Flowers, Oshima y Raju, 1999; Mellenbergh, 1995; Raju, van der Linden y Fleer, 1995; Thissen *et al.*, 1988), generalizaciones del procedimiento de Mantel-Haenszel (Zwick y Thayer, 1996), del procedimiento de estandarización (Dorans y

Schmitt, 1991), extensiones del método de la regresión logística (Agresti, 1990; French y Miller, 1996), métodos basados en el AFC (Oort, 1992) y el procedimiento SIBTEST (Chang, Mazzeo y Roussos, 1996).

Potenza y Dorans (1995) proponen una taxonomía de los estadísticos de detección de DIF en ítems politómicos que clasifica los procedimientos de acuerdo con dos dimensiones: el criterio de equiparación de los grupos y la forma (paramétrica-no paramétrica) en que se estima el funcionamiento del ítem en cada nivel del rasgo. La naturaleza de la estimación del rasgo medido utilizado como variable de equiparación da lugar a métodos basados en la puntuación observada y métodos basados en la variable latente. En los primeros simplemente se utiliza la puntuación total del test como estimación del rasgo latente y en los segundos se utiliza una estimación del rasgo latente, que se obtiene con métodos basados en la TRI (ver Baker, 1992; Hambleton, Swaminathan y Rogers, 1991) o mediante la estimación de la puntuación verdadera con la TCT (ver Lord y Novick, 1968).

La relación entre la puntuación en el ítem y la variable o criterio de equiparación puede ser paramétrica o no paramétrica. Los procedimientos paramétricos estiman el funcionamiento del ítem a cada nivel del rasgo mediante una función matemática, la Curva Característica del Ítem (CCI) utilizada en TRI. Así, se consideran las diferencias de forma de las CCIs entre los grupos como una indicación de que el funcionamiento esperado del ítem en cada nivel del rasgo medido es diferente en los dos grupos y eso significa DIF. Estos procedimientos se consideran paramétricos porque la forma de las CCI se determinan por uno o más parámetros de la función matemática. En contraposición, los procedimientos no paramétricos no utilizan ningún modelo matemático para determinar el funcionamiento

del ítem en cada nivel del rasgo, simplemente consideran el funcionamiento observado del ítem para cada grupo en cada nivel del rasgo. Si hay diferencias entre los grupos en el funcionamiento observado del ítem se considera un indicio que sugiere la existencia de DIF.

En la Tabla 1 se organizan los procedimientos de detección en ítems politómicos de acuerdo a los criterios expuestos. Clasificaciones similares así como una explicación detallada de estos procedimientos –que excede los propósitos de este texto- se pueden encontrar en Hidalgo y Gómez (1999), Millsap y Everson (1993), Penfield y Lam (2000) y Potenza y Dorans (1995).

Tabla 1. *Clasificación de las técnicas de DIF en ítems politómicos*.

| | | Forma de caracterizar el funcionamiento del ítem | | |
|---|---|---|---|---|
| | | Paramétrico | | No paramétrico |
| Forma de estimar el rasgo medido | Puntuación observada | **Regresión Logística Politómica** (Agresti, 1990; French y Miller, 1996).<br>**Análisis discriminante logístico**: (Miller y Spray, 1993) | | **Mantel** (Mantel, 1963).<br>**Mantel-Haenszel Generalizado** (Somes, 1996).<br>**Estandarización** (Dorans y Schmitt, 1991).<br>**Pruebas Z**: (Welch y Hoover, 1993). |
| | Variable latente | Basados en la TRI | **Medidas de área** (Cohen, Kim y Baker, 1993)<br>**Estadístico de Lord** (Cohen, Kim y Baker, 1993)<br>**Test de razón de verosimilitud** (Kim y Cohen, 1998)<br>**Procedimiento DFIT** (Flowers, Oshima y Raju, 1999) | **Polytomous SIBTEST** (Chang, Mazzeo y Roussos, 1996). |
| | | No basados en la TRI | **Análisis factorial confirmatorio** (Oort, 1992) | |

En líneas generales, las ventajas e inconvenientes de estos procedimientos están relacionados con su poder de detección, complejidad computacional, tasa de error tipo I y capacidad para detectar DIF uniforme y no uniforme.

Los procedimientos no paramétricos que se basan en puntuaciones observadas para estimar el rasgo medido (Mantel, Mantel-Haenszel generalizado, estandarización y pruebas Z) presentan la ventaja de ser computacionalmente simples y de tener alto poder de detección de DIF con bajas tasas de error Tipo I cuando las medias de los grupos son similares y el DIF es uniforme. Su desventaja radica en el incremento del error de Tipo I en el caso de grupos con diferente media, que se agudiza más en ítems con mayor discriminación y cuando es menor la fiabilidad de la variable criterio. Su poder de detección de DIF decrece en el caso de los ítems con DIF no uniforme. (Chang, Mazzeo y Roussos, 1996; Welch y Hoover, 1993).

A diferencia de los procedimientos anteriores, el SIBTEST (método no paramétrico que se basa en puntuaciones latentes) tiene buenas tasas de error Tipo I cuando las medias de los dos grupos difieren en diferentes niveles de discriminación del ítem y el tamaño muestral de los grupos es distinto, aunque en estos casos disminuye su poder de detección de DIF y no es apropiado para DIF no uniforme (Chang *et al*., 1996; Zwick, Thayer y Mazzeo, 1997).

Los procedimientos paramétricos basados en la puntuación observada –regresión logística politómica y análisis discriminante logístico- tienen la ventaja de ser eficientes en la detección de DIF no uniforme, pero la desventaja de requerir muestras muy amplias para

realizar una adecuada estimación de los parámetros (French y Miller, 1996; Miller y Spray, 1993).

Los métodos basados en la TRI también necesitan muestras muy grandes, además de tener unos supuestos muy restrictivos en el ajuste del modelo, lo que constituye el principal inconveniente de su aplicación. Por el contrario, son procedimientos bastante potentes en la detección del DIF. De ellos, el método basado en el estadístico de razón de verosimilitud es el más ampliamente utilizado en ítems politómicos (Penfield y Camilli, 2007). Frente a las medidas de área y el estadístico de Lord, tiene las ventajas de no utilizar las matrices de varianzas-covarianzas entre los parámetros estimados para un mismo ítem -que en ocasiones no es muy precisa (Thissen *et al.*, 1988)- y que no es necesario igualar los parámetros, dado que los parámetros de los grupos a comparar se estiman conjuntamente (Kim y Cohen, 1995). En cuanto al procedimiento DFIT propuesto por Raju *et al*. (1995) es un enfoque más novedoso que tiene la ventaja de diferenciar entre DIF acumulativo y no acumulativo, además de proporcionar una medida global del funcionamiento diferencial del test.

El AFC es de los pocos métodos que no necesitó una adaptación al caso politómico. Además, este procedimiento no necesita muestras muy numerosas para su aplicación (en comparación a los procedimientos TRI), y permite comparar más de dos grupos con comodidad.

La unidad de análisis del funcionamiento diferencial del ítem es, como su propio nombre indica, el ítem. No obstante, algunas medidas de DIF incluyen medidas para el Funcionamiento Diferencial del Test (Differential Test Functioning, DTF). Éstas resultan

de indudable interés en esta investigación, ya que un test puede contener algún ítem que presente DIF y, sin embargo, no presentar funcionamiento diferencial a nivel de escala (Camilli, 1993; Zumbo, 2003). Esto fue lo que observó Drasgow en un estudio (Drasgow, 1987) en el que identificó varios ítems que presentaban DIF en relación a las variables sexo y raza en un test de matemáticas; sin embargo, las Curvas Características del Test (CCT) analizadas no identificaban diferencias entre estos grupos al considerar el test en su conjunto. Drasgow argumentó que los ítems con DIF no causaron DTF probablemente porque se compensaron unos con otros al analizar el test completo.

## 3. EQUIVALENCIA DE MEDIDA Y VALIDEZ

La validez de un test se refiere al grado en que las puntuaciones de un test miden lo que pretenden medir. Es, por tanto, el grado en que la evidencia empírica y el razonamiento teórico apoyan la adecuación e idoneidad de las interpretaciones basadas en las puntuaciones de acuerdo con los usos propuestos por el test (Messick, 1989; Prieto y Delgado, 2010).

La concepción teórica de validez ha evolucionado gradualmente a lo largo de los años (Anastasi, 1986; Angoff, 1988). Las sucesivas ediciones de los Estándares de los Tests Psicológicos y Educativos (en lo sucesivo, estándares) publicados en 1954, 1966, 1974, 1985 y 1999 sirven como puntos de referencia, al modificar en cada una de sus versiones -en consonancia con la literatura psicométrica del momento- el tratamiento que se ha dado a este criterio métrico de calidad (un resumen de estos cambios en los estándares puede encontrarse en Kane, 2001 y Alvarado y Santisteban, 2006). Entender

esta evolución histórica del concepto de validez puede ser importante para comprender la importancia que ha adquirido este criterio psicométrico de calidad hasta llegar a convertirse en central.

En un principio lo fundamental era la predicción de un criterio específico, y esa era la utilización principal que se hacía de los tests entre 1920 y 1950 (Kane, 2006). Un ejemplo de los planteamientos teóricos predominantes sobre validez de esa época es, en palabras de Guilford (1946), que "en un sentido muy general, un test es válido para cualquier cosa que correlacione con él" (pag. 429). Después, la concepción predominante era que había un determinado número de tipos de validez, lo que dio lugar a la división tripartita de la validez, utilizándose como vías esenciales para recoger datos en el proceso de validación de los tests el análisis de los contenidos de las pruebas, las correlaciones test-criterio y la entidad de los constructos (Muñiz, 2004). Se trataba, por tanto, de tres tipos de validez: validez de contenido, validez relativa a criterio (predictiva y concurrente) y validez de constructo. Esta visión tripartita no se romperá oficialmente hasta la publicación de los estándares de 1985 (Elosua, 2003).

A partir de entonces se hace hincapié en el significado o interpretación de la medida o puntuaciones de los tests, incrementando el énfasis sobre la validez de constructo como la esencia de una concepción unitaria de validez. La validez de constructo, subsume a la validez de contenido y criterio considerándose el principal modo de validación (Anastasi, 1986; Barbero, Vila y Holgado, 2010; Embretson, 1983; Loevinger, 1957; Messick, 1975, 1980).

Según Messick (1989) hay dos aspectos fundamentales en esta evolución. Uno de ellos es el cambio del énfasis de numerosas evidencias de validez específicas de criterio a un pequeño número de tipos de validez y, finalmente, a una concepción unitaria de la validez. El otro es el cambio de la predicción a la explicación como foco fundamental de la validez, en el sentido de que la utilidad, la relevancia y la importancia de la predicción no pueden utilizarse en la ausencia de la interpretación de las puntuaciones en las que la predicción está basada. En este sentido "la validez es un juicio evaluativo integrado del grado en que la evidencia empírica y las teorías racionales apoyan que las inferencias y acciones basadas en los tests u otros modos de evaluación son apropiados y adecuados" (Messick, 1989, pag. 13).

Actualmente se considera la validez como un proceso continuo, ya que las evidencias se van acumulando, y como éstas siempre son incompletas, nunca se puede dar por finalizado el proceso. Aunque hay muchos caminos de acumulación de evidencias para una inferencia particular esos caminos son esencialmente los métodos de la ciencia (Zumbo, 2007). Las inferencias son hipótesis y su validez es la contrastación de esas hipótesis. En este sentido, el proceso de validación se considera, nada más y nada menos que un proceso de contrastación de hipótesis (Landy, 1986).

Esta concepción de la validez como proceso dinámico y abierto, condicionada a la interpretación de las puntuaciones en relación al uso específico que se haga de ellas, tiene como consecuencia que las fuentes de validación sean múltiples y su importancia varíe en función de los objetivos. Según los estándares (AERA, APA y NCME, 1999) las principales evidencias de validación son: el contenido del test, los procesos de respuesta, la estructura interna de la prueba, las relaciones con otras variables y las consecuencias

derivadas del uso para el que se proponen. Entre las consecuencias se incluye la varianza irrelevante de constructo que puede dar lugar a falta de equivalencia (Messick, 1988).

Garantizar la equivalencia de medida entre grupos aporta evidencias de validez (Penfield, 2005, 2010; Zieky, 2006). Según la clasificación de evidencias de los estándares, el trabajo empírico que aquí se presenta se enmarca en la validación de la estructura interna del test, en el que se ubican la evaluación de la dimensionalidad de la prueba, así como el funcionamiento diferencial del ítem y del test. Así pues, cuando se afirma que las puntuaciones de un test son válidas, a nivel de invarianza lo que realmente estamos diciendo es que la puntuación obtenida tiene un significado específico, asumiendo que este significado es el mismo en los distintos grupos para los cuales el test ha sido validado. En función del uso que se haga de las puntuaciones del test, también será pertinente evaluar las posibles consecuencias de esta utilización como parte del proceso de validación. En particular, será de vital importancia analizar y justificar las consecuencias cuando el test se vaya a emplear para tomar decisiones importantes para las personas, como en el caso de oposiciones, selección de personal, promoción profesional, pruebas de selectividad, permiso de conducir o permiso de armas, entre otros.

El análisis de la equivalencia de medida de un test, por tanto, es parte sustancial del análisis de la validación de las puntuaciones al aplicar el instrumento de medida en cuestión. Para asegurar la equidad de las puntuaciones de sujetos que pertenecen a distintos grupos, éstas tienen que depender únicamente del nivel del sujeto en el constructo medido. Los ítems sesgados crean una distorsión en los resultados del test para los miembros de un grupo particular, de tal modo que sujetos que pertenecen a grupos distintos, aun teniendo el mismo nivel en el constructo medido, obtienen puntuaciones diferentes en dichos ítems;

ello no se debe a un error aleatorio de medida sino a un error sistemático del instrumento de medida, por el que un subgrupo de la muestra resulta beneficiado y otro perjudicado al evaluarles con los ítems en cuestión (Camilli y Shepard, 1994; Gómez y Navas, 1998).

Considerando las implicaciones personales y sociales que puede tener un test, la validación de las puntuaciones de un test es un proceso necesario para interpretar de manera correcta las puntuaciones que se obtengan con él. En este sentido, los estándares (AERA, APA y NCME, 1999) consideran que los constructores de un test deben asumir una responsabilidad y elaborar ítems que estén libres de DIF y DTF en diferentes grupos como género, etnia, o nivel socioeconómico.

# 4. EQUIVALENCIA DE MEDIDA DE UNA PRUEBA CON MODELOS DE ECUACIONES ESTRUCTURALES: EL AFC MULTIGRUPO

## 4.1. MODELOS DE ECUACIONES ESTRUCTURALES

En los últimos 30 años los Modelos de Ecuaciones Estructurales (Structural Equation Modelling, SEM) han llegado a ser una de las más importantes técnicas de análisis de datos en las Ciencias Sociales. De hecho, según algunos autores como Kaplan (2000) se ha convertido en un lenguaje para formular teorías en ciencias sociales y hablar sobre las relaciones entre variables. El surgimiento de estos modelos se debe a dos tradiciones: el análisis factorial desarrollado en el campo de la psicología y el modelado de ecuaciones simultáneas desarrollado en economía y genética.

El origen de los modelos de ecuaciones estructurales data de 1970, año en que el económetra Arthur Goldberger organizó una conferencia sobre modelos que analizaban relaciones causales, a la que invitó a estadísticos, psicómetras, económetras, biómetras y sociómetras. En ella se planteó que no sólo tenia interés estudiar la relación entre variables observables y latentes, sino también entre las propias variables latentes. En esta conferencia fue donde Jöreskog (1973) presentó la primera formulación del Covariance Structure Analysis (CSA) para estimar un los parámetros en un sistema de ecuaciones estructurales lineales, el cual llegó a ser conocido más tarde como LISREL (LInear Structural RELations). Según Mulaik (1986), la importancia del estudio de Jöreskog radica en que unificó análisis factorial, análisis de estructuras de covarianza y modelos de ecuaciones estructurales lineales, en un modelo general único que respaldó, junto a Sörbom, con su famoso programa LISREL (Jöreskog y Sörbom, 1979).

Los SEM pueden definirse como un conjunto de procedimientos que representan hipótesis sobre las medias, varianzas y covarianzas de los datos observados, en términos de un número pequeño de parámetros definidos por un modelo subyacente hipotetizado. Estos modelos engloban y extienden los procedimientos de regresión, el análisis econométrico y el análisis factorial (Bollen, 1989).

En los modelos de ecuaciones estructurales hay una serie de etapas orientadas a minimizar la diferencia entre las covarianzas muestrales y las covarianzas predichas por el modelo propuesto. SEM trata de modelizar la matriz de varianzas-covarianzas de las variables observadas. Para ello, asume que la matriz de covarianzas poblacional de las variables observadas depende de un vector de parámetros a estimar:

$$\sum = \sum(\theta) \qquad (2)$$

donde $\sum$ denota la matriz de covarianzas poblacional de variables observadas, $\theta$ es un vector que contiene los parámetros del modelo y $\sum(\theta)$ es la matriz de covarianzas escrita como una función de $\theta$.

Si el modelo es absolutamente correcto y se conocen todos los parámetros, $\sum$ es exactamente igual a $\sum(\theta)$. En la práctica, los parámetros del modelo se desconocen, por lo que se utiliza una matriz de covarianzas muestral (S) como estimación no sesgada de $\sum$ y se estima el vector $\theta$. Esto último se consigue minimizando alguna función de discrepancia $F[S, \sum(\theta)]$, a partir de la cual se establecen índices de ajuste que permiten evaluar la bondad de ajuste del modelo evaluado (Gómez, 1996).

Hay varias etapas en su realización: (1) especificación del modelo, (2) identificación, (3) estimación, (4) evaluación del ajuste del modelo y (5) reespecificación. En primer lugar se realiza una representación mediante un diagrama de flujos (path diagram) del modelo teórico. Se selecciona una muestra adecuada para los propósitos de la investigación y se recogen los datos. Después, en la etapa de identificación, se comprueba que el modelo sea estimable, esto es, que los parámetros del modelo se puedan derivar a partir de las varianzas y las covarianzas entre las variables observables (ver, por ejemplo, MacCallum, 1995 para una explicación detallada del concepto de identificación en los modelos de ecuaciones estructurales). Se elige el método de estimación más apropiado y, una vez estimado el modelo se procede a la evaluación del ajuste de los datos al modelo especificado. En caso de que el ajuste no sea apropiado, es posible la modificación del modelo, lo que conllevaría un nuevo proceso de identificación y estimación (obviamente, también habría que asegurarse que la modificación llevada a cabo es congruente con el

modelo teórico planteado). Si el ajuste es adecuado el modelo está preparado para ser utilizado.

(1) En la **fase de especificación** se plantea formalmente el modelo, formulando una serie de hipótesis sobre las relaciones entre un conjunto de variables. Estas variables pueden ser observables (medibles directamente) o latentes (constructos no medibles directamente; endógenas (si reciben una influencia direccional de otra variable del modelo) y exógenas (si no la reciben). La relación que se establece entre las variables puede ser direccional o no direccional. Si la relación se define como direccional da lugar a un coeficiente de regresión lineal, y si la relación se define como no direccional da lugar a valores de covarianza entre las variables. También es necesario establecer el valor de los parámetros, que puede ser fijo (si se especifica su valor de antemano) o libre (si se estima su valor a partir del análisis de datos).

(2) En la **fase de identificación** del modelo se pone en correspondencia la información que debe obtenerse (parámetros libres) con la información disponible (matriz varianzas-covarianzas observada) comprobando si hay un único conjunto de parámetros consistente con los datos. Si se encuentra una única solución el modelo se considera identificado. Si, por el contrario, el modelo no puede ser identificado, los parámetros están sujetos a arbitrariedades, de modo que diferentes valores de los parámetros definen el mismo modelo. En este caso, no es posible realizar estimaciones consistentes para todos los parámetros y el modelo no puede ser evaluado empíricamente. (ver, por ejemplo, MacCallum, 1995). Si el modelo está sobreidentificado (el número de parámetros a estimar es menor que el número de varianzas y covarianzas de la matriz

de datos), los grados de libertad son positivos, por lo que el modelo puede ser rechazado y por tanto puesto a prueba.

(3) En la fase de **estimación de los parámetros** se estiman los parámetros libres, mediante métodos iterativos capaces de generar una matriz de varianzas-covarianzas lo más parecida posible a la matriz de varianzas-covarianzas obtenida (S) a partir de los datos utilizados. Los métodos de estimación más utilizados son máxima verosimilitud (ML), mínimos cuadrados generalizados (GLS), mínimos cuadrados ponderados (WLS) y mínimos cuadrados no ponderados (ULS).

(4) En la **fase de ajuste** del modelo se comprueba el grado en que coinciden las matrices S y $\sum$ para determinar si el modelo es correcto y sirve como aproximación al fenómeno real.

Dado que no existe una única medida aceptada para determinar la bondad de ajuste (Ávalo, Lévy, Rial y Valera, 2006), la mayoría de autores abogan por un uso conjunto de varios índices globales en la evaluación de dicho ajuste (Hoyle, 1995; Marsh, Balla, y McDonald, 1988; Tanaka, 1993; Tomás y Oliver, 2004).

Los índices de ajuste pueden dividirse en dos clases: absolutos e incrementales (Hu y Bentler, 1999). Los índices de ajuste absoluto expresan el grado de exactitud en que el modelo global predice satisfactoriamente la matriz de covarianzas observada. Por su parte, las medidas de ajuste incremental comparan el modelo analizado con un modelo de base habitualmente denominado modelo nulo. A menudo, el modelo nulo corresponde al modelo especificado sin ninguna relación entre las variables. Una

revisión más detallada de los índices de bondad de ajuste puede encontrarse en Batista y Coenders (2000), Lévy y Varela (2006) o Tanaka (1993), entre otros.

Los índices de bondad de ajuste absolutos más utilizados son el estadístico $\chi^2$ de bondad de ajuste, el índice de bondad del ajuste (Goodness of Fit Index, GFI) y el error cuadrático medio de aproximación (Root Mean Square Error of Approximation, RMSEA).

El índice absoluto más conocido es el estadístico $\chi^2$ de bondad de ajuste, que sigue una distribución $\chi^2$ con los mismos grados de libertad g que el modelo. La hipótesis nula a contrastar es que el modelo es correcto, y cuanto mayor sea el valor obtenido del estadístico $\chi^2$ en comparación con los grados de libertad, peor será el ajuste (Bollen, 1989). El problema del estadístico $\chi^2$ es que tiende a sobreestimarse cuando el tamaño muestral es grande (Byrne, 1994; 1998), por lo que en estos casos se hace necesario utilizar otros índices para la interpretación del ajuste del modelo.

El índice GFI es una transformación monótona del estadístico $\chi^2$. Su valor está comprendido entre 0 y 1, indicando este último un ajuste perfeto. Un ajuste aceptable tendría un índice próximo a 0,90 (Jöreskog y Sörbom, 1990).

El índice RMSEA representa la bondad del ajuste que podría esperarse si el modelo fuera estimado con la población y no sólo con la muestra extraída de la estimación. Valores de hasta 0’05 indican buen ajuste, valores de hasta 0’08 representan errores de aproximación razonables y valores superiores a 0,1 indican una mala aproximación (Browne y Cudeck, 1993).

Los índices de bondad de ajuste incrementales más utilizados son el índice de ajuste normalizado (Normed Fit Index, NFI), el índice de ajuste no normalizado (Non Normed Fit Index, NNFI o Tucker Lewis Index, TLI) y el índice de ajuste comparativo (Comparative Fit Index, CFI).

El NFI compara la función de ajuste del modelo nulo con la del modelo propuesto (Bentler y Bonnet, 1989). Los valores de este índice varían entre 0 y 1, considerándose aceptables valores superiores a 0,9. Este índice no tiene en cuenta los grados de libertad del modelo propuesto y, a medida que se liberan parámetros, se consiguen modelos más ajustados.

El índice NNFI o TLI es un índice que supera las limitaciones del NFI al considerar los grados de libertad del modelo propuesto y nulo estando, por lo tanto, muy débilmente relacionado con el tamaño muestral. El rango de este índice varía entre 0 y 1, siendo recomendables valores superiores a 0,9.

El índice CFI mide la mejora en la medición de la no centralidad de un modelo (Bentler, 1990). Se trata de una versión revisada del índice de ajuste de Bentler-Bonett (Bentler y Bonett, 1980) que ajusta los grados de libertad y solo adopta valores en el rango de 0 a 1. Aunque en un primer momento se consideró que un valor mayor que 0’90 era representativo de un buen ajuste (Bentler, 1992), revisiones más recientes aconsejan valores cercanos a 0’95 (Hu y Bentler, 1999).

Por otra parte, el índice de validación cruzada esperada (Expected Cross Validation Index, ECVI) se propuso como forma de evaluar, en una muestra simple, la verosimilitud de la validación cruzada realizada en el modelo sobre muestras de similar tamaño de la misma población (Browne y Cudeck, 1989). Específicamente, este índice señala la discrepancia entre la matriz de covarianzas de la muestra analizada y la matriz esperada que se obtendría en otra muestra de tamaño equivalente. La aplicación de ECVI asume una comparación de modelos donde se computa este índice para cada uno de ellos, considerando que el modelo con un valor más pequeño de ECVI exhibirá el mejor potencial para la replicación.

Si el ajuste del modelo es bueno, el modelo teórico propuesto constituirá un reflejo plausible de la realidad y se considerará correcto. Si el ajuste no es bueno, cabe la posibilidad de reespecificar el modelo y volver a ponerlo a prueba.

(5) La **fase de reespecificación** viene guiada fundamentalmente por tres aspectos: (a) el contraste de los multiplicadores de Lagrange (índices de modificación); (b) el contraste de Wald (estadístico *t*) y (c) la matriz de residuos normalizados. Un índice de modificación muestra el decremento mínimo en el valor de $\chi^2$ del modelo si un parámetro fijo se hiciera libre y se volviera a estimar el modelo, por lo que sirve para analizar la multicolinealidad, esto es, buscar indicadores que muestran relaciones significativas con algún factor diferente al especificado inicialmente en el modelo. El estadístico *t* comprueba la significación de los parámetros incluidos en el modelo y el análisis de los residuos normalizados mide la discrepancia entre la matriz de covarianzas estimada y la observada.

Se deben introducir únicamente modificaciones que sean acordes con la teoría y hacerlo de manera secuencial, reexaminado los resultados antes de efectuar la siguiente modificación. En cualquier caso, hay que tener en cuenta que la modificación del modelo se ha basado en los resultados de una muestra concreta. La introducción de modificaciones adecuadas para el ajuste del modelo a la muestra, pero inadecuadas para el ajuste a la población se denomina capitalización del azar (Batista-Foguet, Coenders y Alonso, 2004; MacCallum, Roznowski y Necowitz, 1992).

En relación al marco de trabajo estratégico, Joreskog (1993) distingue entre tres escenarios, que denomina estrictamente confirmatorio, modelos alternativos y generación de modelos.

En el primer caso, el investigador postula un modelo simple basado en la teoría, recoge los datos apropiados y pone a prueba el ajuste del modelo hipotetizado con los datos. Basándose en los resultados el investigador acepta o rechaza el modelo, pero no realiza modificaciones.

En los modelos alternativos, el investigador propone varios modelos alternativos que son congruentes con la teoría. Analizando un conjunto de datos empíricos selecciona el modelo más apropiado para representar los datos.

Por último, la generación de modelos representa el caso donde el investigador, habiendo postulado y rechazado un modelo derivado teóricamente debido a su pobre ajuste a los datos muestrales, procede de manera exploratoria (más que confirmatoria) para modificar y reestimar el modelo. El foco principal de interés, en este caso, es localizar la

fuente de desajuste en el modelo y determinar un modelo que describa mejor los datos muestrales.

Ahora bien, los pros y los contras del ajuste post hoc al modelo han sido debatidos rigurosamente en la literatura. Aunque algunos investigadores han criticado su práctica (por ej. Cliff, 1983; Cudeck y Browne, 1983), otros argumentan que mientras que el investigador sea consciente de la naturaleza exploratoria de sus análisis, el proceso puede ser sustantivamente significativo, porque pueden tomarse en consideración, tanto la significación práctica como la estadística (Byrne, Shavelson y Muthén, 1989; Tanaka y Huba, 1984). Jöreskog (1993), por su parte, considera que si el modelo es rechazado por los datos, el problema es determinar qué está equivocado en el modelo y cómo el modelo debería modificarse para ajustar mejor a los datos.

## 4.2. EL ANÁLISIS FACTORIAL

### 4.2.1. CONCEPTO

El origen de esta técnica se remonta al estudio sobre el patrón de correlaciones de distintas medidas del rendimiento realizado por Spearman a principios del siglo XX (Spearman, 1904). A estas ideas se suman, en las décadas posteriores, las aportaciones de otros investigadores como Thurstone, que populariza el procedimiento con su libro Multiple Factor Analysis (1947), y Lawley (1943) que formula el estimador de máxima verosimilitud.

El Análisis Factorial (AF) es una técnica estadística multivariante que sirve para estudiar la estructura latente o dimensiones que subyacen a las relaciones entre variables, denominados factores o rasgos latentes (Hair, Anderson, Tatham y Black, 1999). Estos factores son inferidos a partir de la puntuación observada o empírica obtenida por cada sujeto tras contestar los ítems de la escala utilizada para evaluar un constructo psicológico en particular (McDonald, 1999, Santisteban, 1990, 2009).

Se distinguen dos formas de AF: Análisis Factorial Exploratorio (AFE) y Análisis Factorial Confirmatorio (AFC). En un AFE el investigador estudia qué estructura factorial se ajusta mejor a los datos sin realizar previsiones sobre el número de factores que subyacen a las relaciones entre variables (que se decide mediante una estrategia empírica), qué variables pesan en cada factor o qué factores correlacionan entre sí. En el AFC, sin embargo, el investigador no solo cuenta con una hipótesis previa acerca de la estructura de las variables latentes, sino que establece a priori el conjunto total de las relaciones entre los elementos que lo configuran, contrastando directamente su modelo teórico (Abad, Olea, Ponsoda y García, 2011).

¿En qué situaciones se debe utilizar una y otra técnica? En palabras de Bollen, "en áreas sustantivas donde aún se conoce poco, el análisis factorial exploratorio puede ser muy valioso ya que permite sugerir patrones subyacentes en los datos. Sin embargo, si existen hipótesis plausibles sobre la estructura de un modelo, entonces el análisis factorial exploratorio puede frustrar las tentativas para probar tales ideas" (Bollen, 1989, p. 228).

El AFC constituye un caso particular de análisis SEM que se ocupa específicamente de los modelos de medida, es decir, de las relaciones entre las variables observadas (ítems

de un test, puntuaciones de un test, calificaciones) y las variables latentes. En un test, el AFC especifica la relación entre las respuestas a los ítems (variables observadas) y el rasgo latente definido por el instrumento de medida (Benson, 1987; Bollen, 1989; Byrne, 1998; Ferrando, 1996a; Muthen, 1984). De esta forma se contrastan los datos presentados con el modelo teórico planteado y mediante índices de bondad de ajuste se evalúa si el modelo es o no acorde con la teoría.

Como todos los procedimientos factoriales analíticos (Floyd y Widaman, 1995; Tinsley y Tinsley, 1987), el AFC asume que un gran número de ítems se utilizan para valorar un pequeño número de variables latentes o constructos. La idea de base del AFC, por tanto, es que para un conjunto de variables observables $X_1$, $X_2$, $X_3$, ..$X_p$ hay una estructura de factores o variables latentes $\xi_1$, $\xi_2$ ..... $\xi_n$ representada en la ecuación factorial siguiente (Jöreskog y Sörbom, 1996):

$$x = \Lambda_x \xi + \delta \qquad (3)$$

donde:

$x$ = vector de $q \times 1$ variables observadas o medidas

$\xi$ = vector de $n \times 1$ factores latentes o variables subyacentes

$\Lambda_x$ = matriz $q \times n$ de cargas factoriales, que relaciona los $n$ factores con las $q$ variables observadas

$\delta$ = vector $q \times 1$ de los errores de medida o residuos de $x$.

Bajo esta metodología, la respuesta observada es una combinación lineal de una variable latente, una carga factorial y un error o residuo. Esta ecuación es el modelo de medida para variables exógenas en el modelado de ecuaciones estructurales. Típicamente,

el vector *x* representa ítems que sirven como indicadores (variables observadas generadas por constructos latentes subyacentes); diferentes ítems sirven como variables indicadoras para diferentes constructos latentes (ξ) en un AFC. Como consecuencia, los coeficientes de regresión o λ que unen los ítems a sus constructos latentes subyacentes son el interés primordial.

### 4.2.2. EL ANÁLISIS FACTORIAL EN VARIABLES ORDINALES

El AFC se basa en los supuestos de normalidad y linealidad. La utilización del AFC en ítems de un test de respuesta dicotómica o politómica supone la violación de este segundo supuesto, ya que los modelos subyacentes son modelos lineales y las relaciones entre los ítems dicotómicos o politómicos no lo son (Byrne, 1998).

En el AFC, la literatura sugiere que cuando los datos ordinales se analizan por el método de estimación de máxima verosimilitud, las estimaciones de los parámetros pueden resultar sesgadas (Rigdon y Ferguson, 1991). Trabajos como los de Johnson y Creech (1983) y Olsson (1979) concluyen que el ajuste del modelo está severamente distorsionado y los parámetros estimados están sesgados cuando se basan en medidas ordinales analizando la matriz de correlaciones de Pearson. Estos errores son menores en algunas condiciones: cuando las variables categóricas se aproximan a una distribución normal, los ítems tienen un alto número de categorías (5 o más) y las variables son simétricas (Atkinson, 1988; Babakus, Ferguson y Jöreskog, 1987; Bentler and Chou, 1987; Muthén y Kaplan, 1985; West, Finch y Curran, 1995).

Aunque hay modelos alternativos para variables ordinales cuando se utilizan como indicadores de variables latentes (Jöreskog, 1990; Mislevy, 1986; Muthén, 1984), habitualmente se utilizan procedimientos diseñados para datos continuos (Muthén y Kaplan, 1985). Breckler (1990) realizó una revisión bibliográfica de artículos que aplicaban modelos de ecuaciones estructurales en los últimos 15 años encontrando que la mayoría de los que empleaban datos tipo Likert habían utilizado como procedimiento de estimación de parámetros Máxima Verosimilitud, cuando su uso no era el más apropiado.

Una solución a este problema consiste en utilizar para el análisis la matriz de correlaciones policóricas en lugar de la matriz de correlaciones de Pearson en las situaciones en las que se asumen variables continuas subyacentes pero los instrumentos de medida con los que se toman los datos son ordinales. Una correlación policórica estima la relación lineal entre dos variables latentes continuas que subyacen a dos variables observadas ordinales que son indicadores manifiestos de aquellas (Flora y Curran, 2004). Un método de extracción apropiado cuando se analiza la matriz de correlaciones policóricas en variables ordinales es el método de mínimos cuadrados ponderados (WLS) y su versión robusta (DWLS). Este procedimiento proporciona errores típicos correctos en muestras grandes (Flora y Curran, 2004; Holgado, Chacon, Barbero y Vila, 2010; Joreskog, 2002). Su versión robusta DWLS se recomienda, además, por sus mayores tasas de convergencia.

## 4.3. EL ANÁLISIS FACTORIAL CONFIRMATORIO COMO PROCEDIMIENTO PARA EVALUAR LA EQUIVALENCIA DE MEDIDA DE UNA PRUEBA EN VARIOS GRUPOS

De los métodos factoriales disponibles, el AFC es actualmente la herramienta más ampliamente utilizada en el estudio de la equivalencia de medida en múltiples grupos (Brown, 2006; Byrne, Shavelson y Muthén, 1989; Jöreskog, 1971; Meredith, 1993; Millsap y Everson, 1993; Reise, Widaman y Pugh, 1993; Steenkamp y Baumgartner, 1998; Vandenberg, 2002; Vandenberg y Lance, 2000; Widaman y Reise, 1997). En estos últimos años, el interés en el estudio de la equivalencia factorial y asuntos relacionados es mayor que en cualquier otro momento de los últimos 100 años (Millsap y Meredith, 2007).

Dado que la relación entre las variables observadas y los constructos subyacentes hipotetizados puede modelizarse mediante AFC según la ecuación 3:

$$x = \Lambda_x \xi + \delta \qquad (3)$$

Y asumiendo que los errores de medida aleatorios tienen un valor esperado igual a cero, la esperanza de *x* en esta ecuación puede ser escrita como:

$$E(x) = \Lambda_x \xi \qquad (4)$$

Por otra parte, asumiendo que los errores de medida aleatorios, no están correlacionados unos con otros ni con los factores subyacentes, la matriz de varianzas-covarianzas para una población dada $\Sigma_x$ puede ser expresada como:

$$\Sigma_x = \Lambda_x \Phi \Lambda'_x + \Theta_\lambda \qquad (5)$$

donde:

$\Lambda'_x$ es la traspuesta de $\Lambda_x$

$\Phi$ es la matriz de varianzas-covarianzas entre los factores ($\xi$)

$\Theta_\lambda$ es una matriz diagonal de las varianzas error.

$\Sigma_x$ representa la matriz de varianzas-covarianzas para una población dada. Cuando hay varias poblaciones, tenemos una matrices $\Sigma_x$, $\Lambda_x$, $\Phi$ y $\Theta_\lambda$ diferentes para cada población.

En un análisis factorial multigrupo, el modelo teórico se compara con la estructura observada en dos o más muestras. Habitualmente se sigue la estrategia de Jöreskog de comparación de estructuras de covarianzas (Jöreskog, 1971; 1993) para comprobar la invarianza de medida. En esta estrategia se organizan modelos anidados en un orden jerárquico, con la disminución sucesiva del número de parámetros (o el aumento de los grados de libertad), lo que implica que se van añadiendo restricciones al modelo, forzando la igualdad de parámetros entre los grupos de manera sucesiva. Estos modelos cada vez más restrictivos se evalúan en términos del ajuste de sus datos al modelo (Cheung y Rensvold, 1999, 2002; Milfont y Fischer, 2010; Steenkamp y Baumgartner, 1998; Vandenberg y Lance, 2000).

Para ello, además de realizar el estudio independiente de cada uno de los modelos con los índices de ajuste ya comentados anteriormente, se evalúa comparativamente el

ajuste de los modelos anidados, calculando la diferencia entre los $\chi^2$ de los modelos ($\Delta\chi^2$). La significación estadística de esta diferencia se determina utilizando la diferencia en grados de libertad ($\Delta$g.l.) a un nivel $\alpha$ especificado a priori. Además del incremento en $\chi^2$, siguiendo los criterios propuestos por Cheung y Rensvold (2002), hay que tener en cuenta también la diferencia entre los valores en el índice comparativo de Bentler (CFI). Si esta diferencia entre dos modelos anidados es superior a 0'01 debería rechazarse el modelo con más restricciones.

Varios autores (p.e., Borges, van den Bergh y Hox, 2001; Byrne, Shavelson y Muthen, 1989; Milfont y Fischer, 2010; Vandenberg y Lance, 2000, Wu, Li y Zumbo, 2007) consideran una clasificación de los distintos modelos que proviene de los modelos de tests (Anderson y Gerbing, 1988) y distingue entre modelos que ponen a prueba aspectos de invarianza de medida (modelos que evalúan invarianza de constructo, cargas factoriales, ordenadas en el origen o interceptos y varianzas error) y modelos que ponen a prueba aspectos de invarianza estructural (modelos que evalúan invarianza de varianzas, covarianzas, y medias de las variables latentes).

Los modelos de invarianza de medida son modelos que ponen a prueba las relaciones entre las variables medidas y los constructos latentes. Son cuatro: invarianza de configuración, métrica, escalar y de varianza error. Los modelos de invarianza estructural son modelos que se refieren únicamente a las variables latentes y son tres: invarianza de las varianzas de los factores, de las covarianzas de los factores y de las medias de los factores. Conviene aclarar aquí que el término de invarianza estructural utilizado en este contexto tiene un significado diferente al de la literatura transcultural, en el que se evalúa si los indicadores están relacionados con el constructo de forma no trivial (Fontaine, 2005).

A continuación de la Tabla 2 se presenta una descripción de los distintos modelos.

Tabla 2. *Modelos de invarianza*

| MODELO | | NOTACIÓN SIMBÓLICA* | HIPÓTESIS CONTRASTADA | SIGNIFICADO CONCEPTUAL DE LA HIPÓTESIS |
|---|---|---|---|---|
| Pruebas sobre invarianza de medida | | | | |
| **Modelo 1** | Invarianza de configuración | $\Lambda^{g}_{form} = \Lambda^{g'}_{form}$ | Misma estructura factorial en ambos grupos | Ambos grupos asocian los mismos subconjuntos de ítems con los mismos constructos (el dominio cognitivo es el mismo). |
| **Modelo 2** | Invarianza métrica | $\Lambda^{g} = \Lambda^{g'}$ | Igualdad de cargas factoriales | La fuerza de las relaciones entre cada ítem y su constructo subyacente es la misma en ambos grupos. |
| **Modelo 3** | Invarianza escalar | $\tau^{g} = \tau^{g'}$ | Igualdad de interceptos | Las diferencias entre grupos que indican los ítems son las mismas en todos los ítems. |
| **Modelo 4** | Invarianza de las varianzas error | $\Theta^{g}_{\lambda} = \Theta^{g'}_{\lambda}$ | Igualdad de varianzas error | Los ítems tienen la misma consistencia interna en ambos grupos. |
| Pruebas sobre invarianza estructural | | | | |
| **Modelo 5** | Invarianza de la varianza de los factores | $\Phi_{j}^{g} = \Phi_{j}^{g'}$ | Igualdad de las varianzas de los factores | La variabilidad con respecto a los constructos es la misma en ambos grupos. |
| **Modelo 6** | Invarianza de la covarianza de los factores | $\Phi_{jj'}^{g} = \Phi_{jj'}^{g'}$ | Igualdad de las convarianzas entre los factores | Las relaciones entre los constructos (p.e. correlaciones) son las mismas en ambos grupos |
| **Modelo 7** | Invarianza de las medias latentes | $\kappa^{g} = \kappa^{g'}$ | Igualdad de medias | La media de cada constructo es la misma en ambos grupos. |

*Nota: Los superíndices g y g' indican dos grupos distintos. Para abreviar, se muestra únicamente el caso de dos grupos, pero cada hipótesis se puede generalizar a K grupos.

**Modelo 1. Invarianza de configuración.**

En este modelo se pone a prueba la hipótesis nula de que existe el mismo patrón de factores de carga fijos y libres en cada grupo (Horn y McArdle, 1992), por lo que para ponerlo a prueba ambos grupos se analizan simultáneamente, dejando invariante el patrón de cargas factoriales. En otras palabras, cada ítem tiene que pertenecer al mismo factor en todos los grupos, pero se permite que todos los parámetros estimados varíen entre los dos

grupos. Por tanto, los índices de ajuste de este modelo base de igualdad de patrones factoriales reflejan el ajuste de los parámetros de los ítems estimados separadamente para cada grupo. La cuestión por tanto, es saber si la matrices $\Lambda_x$, $\Phi$ y $\Theta_\lambda$ son equivalentes en los grupos, lo que presupone que el número de factores subyacentes y los patrones de los factores son los mismos en las poblaciones o grupos objeto de estudio.

Si se da esta equivalencia, el número de factores y el patrón de matrices de cargas factoriales es similar entre los grupos, por lo que podemos decir que los ítems definen los mismos factores, aunque los pesos de los ítems sobre las escalas pueden variar entre grupos. Dicho de otro modo, aceptar esta hipótesis de equivalencia implica que los grupos asocian los mismos ítems con los mismos constructos (Coenders, Batista-Foguet y Saris, 2005; Meredith, 1993; Riordan y Vandenberg, 1994).

Este modelo se considera el modelo base con el que se evalúan los modelos de invarianza posteriores.

**Modelo 2. Invarianza métrica.**

En este modelo, además del cumplimiento de la invarianza de configuración, se requiere que las saturaciones factoriales sean iguales entre grupos. Por tanto, no solo la composición de los factores debe ser constante, sino también el peso de cada variable en la constitución de cada factor.

En la invarianza métrica se pone a prueba la hipótesis nula de que las cargas factoriales para cada ítem son invariantes en los grupos ($\Lambda^g = \Lambda^{g'}$), poniendo así a prueba

un modelo que añade la restricción de las cargas factoriales que ahora obligatoriamente tienen que ser iguales en los dos grupos (sirve para identificar DIF no uniforme). Si se da esta equivalencia, la Ecuación 3 será idéntica en las dos poblaciones, o lo que es lo mismo, dos personas de diferentes poblaciones con un vector idéntico de puntuaciones factoriales ($\xi$) tendrán el mismo vector de puntuaciones esperadas.

Aceptar esta hipótesis de equivalencia implica que la fuerza de la relación entre cada ítem y su constructo o factor subyacente es idéntica entre los grupos. En este caso, puede decirse que los factores significan lo mismo en los grupos y es legítimo comparar a los grupos en las varianzas (covarianzas) de los factores latentes incluidos en el modelo.

Este modelo de invarianza de medida debe establecerse al menos en parte para que las posteriores pruebas sean significativas. La invarianza métrica junto con la invarianza configural son las dos pruebas de invarianza más utilizadas en la literatura.

**Modelo 3. Invarianza escalar**

Este modelo pone a prueba la hipótesis nula de que las ordenadas en el origen de las ecuaciones de regresión de los ítems sobre las variables latentes no varían en los grupos ($\tau^{g} = \tau^{g'}$), aunque puede haber diferencias en las medias de los factores. La invarianza escalar incluye la igualdad de las cargas factoriales y de los interceptos conjuntamente (Batista y Coenders, 2000; Hui y Triandis, 1985).

Este modelo contrasta si las diferencias de medias entre los grupos en las variables se explican por las diferencias de medias en los factores latentes, por lo que al aceptar esta

hipótesis de equivalencia se considera probada la invarianza de medida, en el sentido de no hay DIF de ningún tipo. Tras comprobar la invarianza escalar (basta con que se cumpla para una parte de los ítems) se pueden comparar las medias de los grupos en los factores (Abad *et al*., 2011; Chan, 2000; Little, 1997).

**Modelo 4. Invarianza de las varianzas residuales o error**

Este modelo requiere que haya la misma cantidad de error de medida de cada ítem en ambos grupos. Por lo que pone a prueba la hipótesis nula de que las varianzas error sobre las variables latentes de cada ítem son iguales en los grupos.

La cuestión de si existe o no igualdad de varianzas de los términos de unicidad/error (matrices $\Theta_\lambda$) entre los grupos puede proporcionar información útil respecto a la fiabilidad de los instrumentos de medida. Si se satisface la condición de igualdad de matrices $\Lambda_x$ y $\Phi$ (estrictamente hablando, solo la igualdad de $\Lambda_x$ y de las varianzas de los factores) en las distintas poblaciones, la igualdad de las matrices $\Theta_\lambda$ implicaría la igualdad de la fiabilidad de las variables medidas en las poblaciones, lo que Byrne denomina equivalencia de fiabilidad (Byrne, 1998).

**Modelo 5. Invarianza de la varianza de los factores**

Consiste en poner a prueba la hipótesis nula de que las varianzas de los factores son invariantes en los grupos ($\Phi_j^g = \Phi_j^{g'}$). En ocasiones se utiliza como un complemento del modelo de invarianza métrica, donde las diferencias en las varianzas de los factores se interpretan como un reflejo de las diferencias en los grupos en la calibración de las puntuaciones verdaderas (ver por ejemplo, Schaubroeck y Green, 1989; Schmitt, 1982).

**Modelo 6. Invarianza de las covarianzas entre los factores**

Consiste en poner a prueba la hipótesis nula de igualdad de las covarianzas entre los factores en los grupos ($\Phi_{jj'}^{g} = \Phi_{jj'}^{g'}$). Se utiliza en ocasiones como un complemento del modelo de invarianza de configuración, donde las diferencias en las covarianzas entre los factores se interpretan como un reflejo de las diferencias en las asociaciones conceptuales entre las puntuaciones verdaderas (Schmitt, 1982). En el caso de aceptar esta hipótesis en combinación con las anteriores, tendríamos que las correlaciones entre los constructos son iguales en los distintos grupos, lo que es una restricción muy fuerte y poco probable, aún en el caso de muestras aleatorias de la misma población (Meredith y Horn, 2001).

**Modelo 7. Invarianza de medias en los grupos**

Consiste en poner a prueba la hipótesis nula de la invarianza factorial de medias en los grupos (p.e. $\kappa^{g} = \kappa^{g'}$), que suele realizarse para probar las diferencias de nivel entre grupos, en el rasgo de interés.

En los modelos de invarianza revisados en la Tabla 2, a excepción del modelo 1, es poco probable encontrar en la práctica que se cumpla la invarianza total de medida (Horn, 1991; Horn, McArdle y Mason, 1983; Steenkamp y Baumgartner, 1998). Para hacer frente al objetivo poco realista y quizá demasiado estricto de que las restricciones de invarianza deben realizarse sobre todos los parámetros y en todos los grupos, Byrne, Shavelson y Muthén (1989) introdujeron el concepto de invarianza parcial de medida, en el que solo se restringe la igualdad de un subconjunto de parámetros en un modelo, mientras que se deja variar libremente entre los grupos al resto. Por lo tanto, la invariancia de medida parcial

puede permitir comparaciones apropiadas entre grupos en los casos en los que no se obtenga la invariancia de medida completa.

La invarianza de medida parcial puede ser evaluada en dos casos: (1) cuando hay invarianza de medida entre algunos, pero no entre todos los grupos y (2) cuando alguno, pero no todos los parámetros, son invariantes entre grupos (Valdenberg y Lance, 2000).

Dado que no hay criterios claros para utilizar la invarianza de medida parcial, Valdenberg y Lance (2000) recomiendan que se establezca la invarianza de configuración completa y la invarianza métrica (al menos parcial) antes de poner a prueba cualquier otro modelo de invarianza parcial. Asimismo, argumentan que la invarianza métrica parcial solo se justifica en el caso de que los parámetros que se dejan variar libremente entre los grupos son una minoría de los indicadores (ver también van de Vijver y Poortinga, 1982). Un ejemplo práctico de invarianza parcial de medida puede consultarse en Milfont, Duckitt y Cameron (2006).

Las denominaciones utilizadas aquí en los diferentes modelos de invarianza no son las únicas que aparecen en la literatura. Otros autores (Elosua, 2005; Meredith, 1993; Meredith y Teresi, 2006; Widaman y Reise, 1997) utilizan los términos invarianza factorial suave, invarianza factorial fuerte, e invarianza factorial estricta para referirse a los modelos 2, 3 y 4, asumiendo un orden jerárquico, de manera que cada modelo incluye las restricciones del modelo anterior.

En cualquier caso, queda reflejado que, cuando se evalúa la existencia de un modelo factorial común entre las poblaciones se pueden realizar varios tests de invarianza.

Existe cierta jerarquía en el desarrollo de estos tests (Byrne, 1998; Joreskög y Sörbom, 1989). Por ejemplo, una prueba sobre la igualdad de matrices $\Theta_\lambda$ sólo debe realizarse si se ha encontrado igualdad de matrices $\Lambda_x$ y para realizar un test de equivalencia estructural o igualdad de matrices $\Phi$ previamente se tiene que probar que las matrices $\Lambda_x$ son comparables entre poblaciones.

Asimismo, sin el requisito de equivalencia de configuración, no tiene demasiado sentido examinar la invarianza de $\Lambda_x$, $\Phi$ y $\Theta_\lambda$ en los grupos. Sin embargo, Byrne, Shavelson y Muthén (1989) encuentran que, dado un suficiente número de variables indicadoras por factor, la equivalencia de las matrices $\Phi$ y $\Theta_\lambda$ puede ser evaluada, aunque dentro del contexto de la invarianza parcial de medida.

Asimismo, debe establecerse la invarianza métrica (al menos parcial) antes de poner a prueba los modelos 3 (invarianza escalar) y 4 (invarianza de las varianzas error).

Tal y como apunta Byrne (1998) puede observarse que la evaluación de la invarianza desde el modelo 2 hasta el modelo 4 se corresponde con la misma secuencia recomendada por Gulliksen y Wilks (1950) para comprobar la homogeneidad de los modelos de regresión en varios grupos.

En realidad, solo los modelos de medida -modelos 1 al 4- se organizan en un orden jerárquico incrementándose sucesivamente los requisitos de igualdad de un modelo al siguiente. En estos modelos, cada test de invarianza realizado resulta más restrictivo que el anterior, y solo se puede utilizar un modelo si se ha encontrado equivalencia entre los

grupos en los modelos previos en el orden jerárquico (Milfont y Fischer, 2010). Por el contrario, los modelos estructurales -modelos 5 al 7- no son jerárquicos o secuenciales.

De los modelos estructurales, con frecuencia, el modelo 6 y el modelo 5 se combinan en una prueba más global de igualdad en los grupos de las matrices de varianzas/covarianzas de las variables latentes ($\Phi^g = \Phi^{g'}$). En caso de haber más de un factor subyacente, la equivalencia de la estructura teórica, representada por la correlación de los factores latentes tiene interés (Byrne, 1998; Vandenberg y Lance, 2000). En consecuencia, ambos, varianzas y covarianzas de los factores pueden ser evaluados para su equivalencia entre los grupos, aunque esta última tiene más interés con respecto a la igualdad de la estructura teórica. En investigaciones sustantivas, en las que interesa probar las diferencias observadas entre medias, es importante saber que ambos, las cargas factoriales del ítem y las relaciones entre los factores latentes, son equivalentes entre los grupos. Además, algunos autores han señalado que es bastante posible que los ítems sean equivalentes entre los grupos aunque las relaciones entre los factores latentes no lo sean (Byrne, 1998; Drasgow y Kanfer, 1985; Meredith, 1964).

Existe consenso en que los modelos estructurales no son una condición necesaria para establecer invarianza de medida, porque la igualdad en esos elementos no está involucrada en la definición de las relaciones entre los ítems y los factores (Little, 1997; Meredith y Millsap, 1992; Millsap, 1998; Widaman y Reise, 1997). De hecho, explicar o predecir las diferencias de grupo en la media de, la varianza de, y las interrelaciones entre los factores son a menudo fruto de una investigación mucho más de fondo (Wu *et al*., 2007). Desafortunadamente, el mismo acuerdo no se ha alcanzado sobre la necesidad de la

igualdad en los cuatro primeros modelos de invarianza (Cheung y Rensvold, 2002; DeShon, 2004; Lubke y Dolan, 2003; Little, 1997; Vandenberg y Lance, 2000).

De manera exhaustiva y basándose en una revisión de los estudios que utilizan procedimientos para comprobar la invarianza, Vandenberg y Lance (2000) consideran que hay ocho pruebas de invarianza y una secuencia determinada entre ellas (ver Figura 1) basándose en fundamentos conceptuales y estadísticos. En esta secuencia comienzan por una prueba global de la igualdad de las matrices de covarianzas en los grupos o, lo que es lo mismo, poner a prueba la hipótesis nula de la invarianza de las matrices de covarianza (p.e. $\Sigma^{g} = \Sigma^{g'}$). Jöreskog y Sörbom (1971, 1989) también recomiendan que este test de igualdad de las matrices de varianzas covarianzas sea realizado antes de cualquier otro test de equivalencia. Cuando no se cumple esta igualdad, otros tests de invarianza deben realizarse para precisar la fuente de la desigualdad. Alternativamente, cuando se da esta igualdad, los grupos pueden tratarse como equivalentes y, por lo tanto, no sería necesario realizar los siguientes tests de invarianza. Los datos de los diferentes grupos deben unirse y todos los análisis restantes tienen que basarse en esta matriz única de varianzas-covarianzas.

Sin embargo, algunos autores han cuestionado la utilidad de este test en particular (Byrne, 1998; Rock, Werts y Flaugher, 1978) basándose en que esta prueba indica que la invarianza es razonablemente sostenible cuando más tests específicos de equivalencia llegan a los mismos resultados; no hay acuerdo, por tanto, en términos de qué hacer cuando se demuestra la igualdad de matrices $\Sigma_x$ en ambos grupos, por lo que este test ya no es un requisito previo para realizar un estudio de invarianza (Byrne, 2001, 2004, 2008).

La secuencia recomendada por Vandenberg y Lance (2000) entre las distintas pruebas para comprobar la equivalencia de la medida entre grupos es la siguiente:

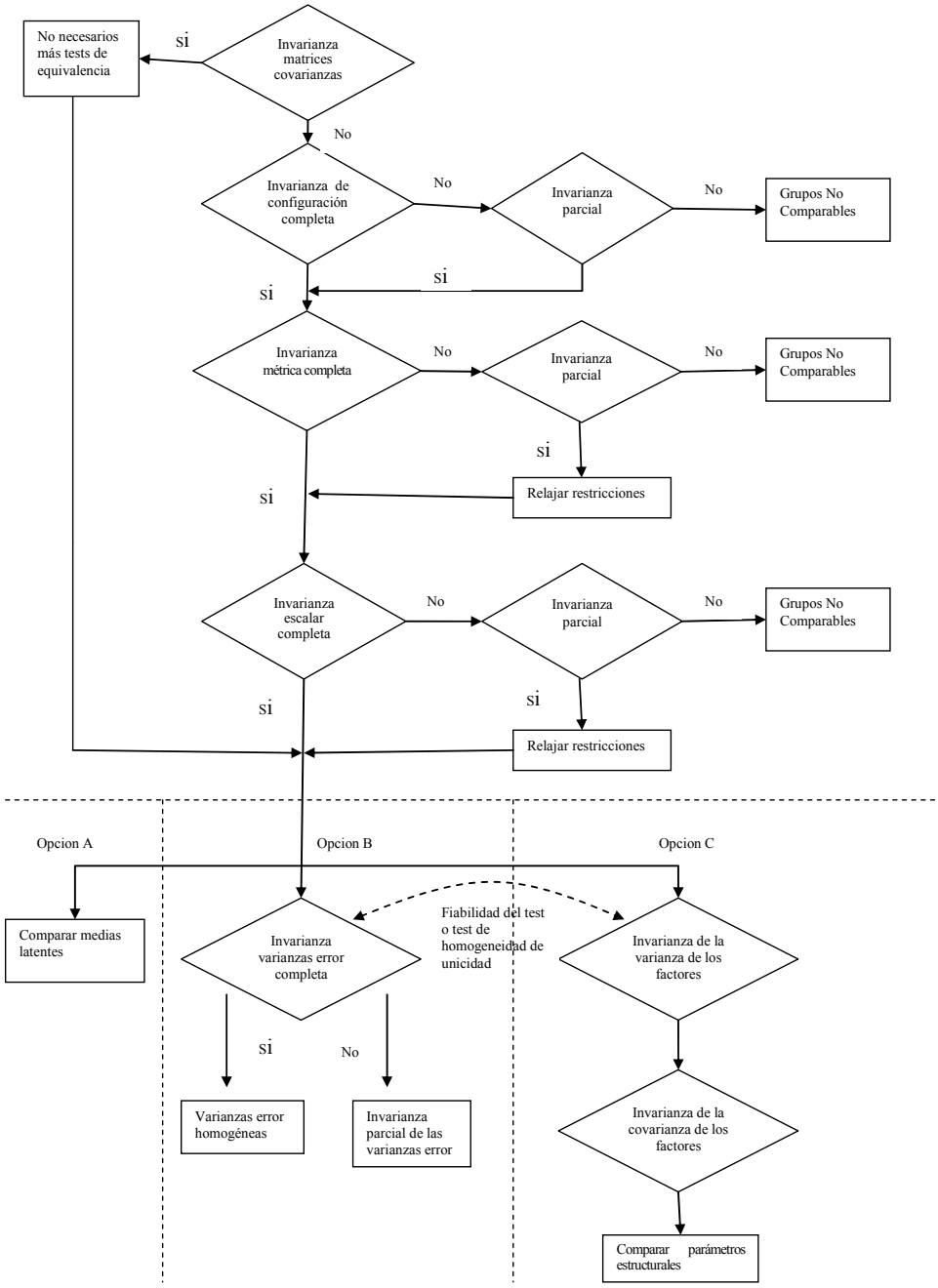



Figura 1. Diagrama de flujo que muestra la secuencia recomendada para comprobar la invarianza/equivalencia de la medida (Vandenberg y Lance, 2000).

La Figura 1 representa de manera exhaustiva todas las comprobaciones que pueden realizarse para poner a prueba la equivalencia de la medida. En la práctica es muy raro

encontrar estudios que hayan realizado los ocho tests de invarianza; habitualmente se eligen las pruebas basándose en las necesidades de la investigación en particular. En general, los tests de equivalencia más frecuentes son los de invarianza de configuración y métrica, aunque, como destacan Vandenberge y Lance (2000) en muchas ocasiones se realizan inferencias incorrectas por no haber realizado ninguna de las pruebas de invarianza de medida.

En la mayoría de ocasiones se excluye de los análisis el estudio de la equivalencia de ordenadas en el origen y de medias. Así, en una escala o subescala unidimensional de *n* ítems con *m* opciones de respuesta para cada ítem, en la que $\xi$ representa el constructo latente, la relación entre este constructo subyacente y la puntuación en el ítem puede expresarse como:

$$x_i = \lambda_i \xi + \delta_i \qquad (6)$$

donde :

$x_i$ es la puntuación observada en el ítem *i*.

$\lambda_i$ es el factor de carga para el ítem *i*.

$\delta_i$ es el término error/residual para el ítem *i*.

Al nivel del ítem la puntuación esperada es:

$$E(x_i) = \lambda_i \xi \qquad (7)$$

Y la varianza:

$$\sigma^2_{xi} = \lambda_i^2 \sigma^2_{\xi} + \sigma^2_{\delta i} \qquad (8)$$

En suma, a nivel de puntuación total en la subescala tendríamos que:

$$E(x) = \lambda \xi \qquad (9)$$

Y:

$$\sigma_x^2 = \lambda^2 \sigma_\xi^2 + \sigma_\delta^2 \qquad (10)$$

Según estas ecuaciones: (a) la relación entre $x_i$ y $\xi$ es lineal (también válido a nivel de escala o subescala); (b) el término error ($\delta_i$) tiene esperanza cero y no está correlacionado con $\xi$ (es aleatorio). Este último supuesto conlleva que la varianza residual a nivel de puntuación total en la subescala sea igual a la suma de las varianzas residuales a nivel de ítem.

En este modelo de AFC, se asume un modelo unidimensional, con lo que las puntuaciones al ítem y a la escala total (o subescala) dependen solo de un constructo latente.

A nivel de ítem la cantidad $\lambda_i \xi$ es la puntuación esperada del ítem o puntuación verdadera. Si dos personas tiene la misma puntuación en la variable latente, entonces tendrán la misma puntuación esperada o puntuación verdadera en el ítem *i* (ecuación 7) o en la subescala (ecuación 9).

La varianza de la puntuación observada de un ítem es igual a la varianza de su puntuación verdadera más la varianza de las puntuaciones error o residuales, a nivel de ítem (ecuación 8) y de puntuación total en la subescala (ecuación 10).

Se puede observar que la relación aditiva entre la varianza de la puntuación verdadera y la varianza error es idéntica a la de la TCT, que establece que la varianza de

las puntuaciones empíricas es igual a la varianza de las puntuaciones verdaderas más la varianza error.

Si se cumple la invarianza métrica (modelo 2) en un AFC de un rasgo unidimensional, esto es, si las cargas factoriales del ítem son invariantes en las dos poblaciones ($\lambda_i^g = \lambda_i^{g'}$), se da este tipo de equivalencia de medida en el test, lo que significa que dos personas, una de cada población, con la misma puntuación en la variable latente, tendrán idénticas puntuaciones verdaderas (o esperadas) a nivel de ítem para todos los ítems (Raju, Laffitte y Byrne, 2002).

Si hay invarianza/equivalencia de la puntuación verdadera a nivel de ítem para todos los ítems, entonces también la habrá a nivel de puntuación total en la subescala o escala. Sin embargo, el caso contrario no es necesariamente cierto. Es teóricamente posible que a nivel de escala las cargas factoriales sean iguales ($\lambda^g = \lambda^{g'}$) y que a nivel de ítem las cargas factoriales no lo sean, ya que debido a su naturaleza aditiva, es posible que las diferencias encontradas a nivel de ítem se cancelen entre sí a nivel de escala.

Una definición más estricta de equivalencia de medida puede requerir que a nivel de ítem las varianzas error sean iguales en las dos poblaciones ($\sigma^2_{\delta ig} = \sigma^2_{\delta ig'}$) para todos los ítems. Cuando esto sucede, y si las varianzas de las puntuaciones factoriales son también iguales, de acuerdo con la TCT (Lord y Novick, 1968), las fiabilidades del ítem serán las mismas en las dos poblaciones. Teniendo en cuenta la ecuación 10 y que la varianza residual a nivel de puntuación total en la subescala es igual a la suma de las varianzas residuales a nivel de ítem, esta definición más estricta de equivalencia de medida requiere

que, tanto las varianzas error como las de las puntuaciones verdaderas a nivel de la subescala sean iguales, por lo que la fiabilidad será la misma en las dos poblaciones.

Habitualmente, desde el AFC se ha focalizado únicamente en el análisis de las estructuras de covarianzas, con lo que se asume que las puntuaciones observadas representan desviaciones de sus medias, por lo que se excluye el estudio de las ordenadas en el origen. Sin embargo, algunos autores no están de acuerdo con esta postura (Chan, 2000; Little, 1997) y consideran que el análisis de equivalencia debe incluir el análisis tanto de estructuras de medias como de covarianzas (Mean and Covariance Structural Analysis, MACS).

En el análisis MACS, en una escala unidimensional de *n* ítems con *m* opciones de respuesta para cada ítem, en la que $\xi$ representa constructo latente, la relación entre este constructo subyacente y la puntuación al ítem puede expresarse como:

$$x_i = \tau_i + \lambda_i \xi + \delta_I \qquad (11)$$

donde :

$x_i$ es la puntuación observada en el ítem *i*.

$\tau_i$ es la ordenada en el origen o intercepto del ítem *i*.

$\lambda_i$ es el factor de carga para el ítem *i*.

$\delta_i$ es el término error/residual para el ítem *i*.

La esperanza de $x_i$ puede expresarse como:

$$E(x_i) = \tau_i + \lambda_i \xi \qquad (12)$$

Y la varianza sería:

$$\sigma^2_{xi} = \lambda_i^2 \sigma^2_{\xi} + \sigma^2_{\delta i} \qquad (13)$$

expresión idéntica a la Ecuación 7.

La ordenada en el origen de la ecuación 12 depende de la media de $x_i$, la media de $\xi$ y $\lambda_i$. De hecho, $\tau_i$ puede expresarse como:

$$\tau_i = \mu_{xi} - \lambda_i \mu_\xi \qquad (14)$$

En esta ecuación, $\tau_i$ es cero cuando la media de $x_i$, y $\xi$ son igual a 0. La ordenada en el origen de una segunda población (denotada como g´) para el mismo ítem puede expresarse como:

$$\tau_i^{g'} = \mu_{xi}^{g'} - \lambda_i^{g'} \mu_\xi^{g'} \qquad (15)$$

Con respecto a los interceptos, como ya se ha visto en el modelo 3, la invarianza escalar se define como igualdad de ordenadas en el origen ($\tau_i^g = \tau_i^{g'}$). La igualdad de las cargas factoriales ($\lambda_i^g = \lambda_i^{g'}$) o de medias ($\mu x_i^g = \mu x_i^{g'}$ y $\mu_\xi^g = \mu_\xi^{g'}$) por sí solas no garantizan la igualdad de interceptos excepto en el caso de que todas las medias sean iguales a cero.

El hecho de que la igualdad de las cargas factoriales no implique necesariamente la igualdad de ordenadas en el origen sugiere que las medias del ítem pueden ser diferentes para las dos poblaciones. Asimismo, la prueba de invarianza de las medias (modelo 7) trata sobre lo que en la literatura sobre DIF se denomina impacto: una diferencia real en el nivel del rasgo de dos grupos.

Según Raju *et al.* (2002) hay que esclarecer si las diferencias estadísticamente significativas de las ordenadas en el origen en dos grupos reflejan DIF o impacto. Por una

parte las ecuaciones 14 y 15 dependen de las medias del ítem y del factor; por lo tanto, las diferencias entre los interceptos del ítem probablemente reflejen impacto. Sin embargo la parte derecha de la igualdad de estas ecuaciones parece implicar que las medias observadas del ítem se modifican por las diferencias en las medias del factor para las dos poblaciones; por lo tanto, una diferencia estadísticamente significativa en las ordenadas en el origen puede reflejar DIF/no equivalencia de medida o una pérdida de ajuste al modelo de medida hipotetizado. Además, la interpretación de una diferencia significativa de interceptos (efecto principal) en la presencia de una diferencia significativa de pendiente (interacción) puede ser problemática.

De hecho, no hay consenso claro sobre la necesidad de evaluar la igualdad de las ordenadas en el origen. Este estudio se alinea con las tesis de autores como Little (1997), Meredith (1993), Steenkamp y Baumgartner (1998) o Coenders, Batista y Saris (2005) que consideran que, además de igualdad de cargas factoriales, tiene que haber igualdad de ordenadas en el origen para poder comparar las medias de los factores entre grupos (aunque bastaría que se cumpliese para una parte de los ítems de cada dimensión). Sin embargo, otros autores como Flowers, Raju y Oshima (2002) consideran que la diferencia de interceptos refleja impacto o que no tiene una interpretación clara, no estando claro si refleja DIF o impacto (Raju *et al*., 2002).

Labouvie y Ruetsch (1995) consideran que solo es necesario que se satisfaga la equivalencia métrica a nivel de escala, esto es, sobre conjuntos de ítems y no sobre cada ítem individualmente. Proponen un método como alternativa conceptual en el contexto del AFC, en el que se imponen restricciones de igualdad sobre cargas factoriales e interceptos a grupos de ítems en vez de a cada ítem de forma individual. Sin embargo, este

procedimiento ha sido duramente criticado. Meredith (1995) considera que es un método exploratorio por lo que no puede servir para contrastar hipótesis y que, al trabajar sobre conjuntos de ítems, la presencia de invarianza parcial de algunos ítems puede dar una impresión errónea sobre las diferencias entre los grupos. Drasgow (1995b) considera que es un método interesante pero que tiene que resolver diversos problemas técnicos.

En todo caso, el AFC multigrupo ha demostrado ser un procedimiento eficaz para el estudio de la invarianza de medida en varios grupos (Brown, 2006; Meredith, 1993; Millsap y Everson, 1993; Millfont y Fischer, 2010; Steenkamp y Baumgartner, 1998; Vandenberg, 2002).

Meade y Lautenschlager (2004b) utilizan datos simulados con diversas diferencias en las cargas factoriales entre dos grupos para valorar la eficacia de las pruebas de equivalencia basados en el AFC multigrupo para detectar distintas pérdidas de equivalencia de medida. Encuentran que los tests de equivalencia discriminan bien el funcionamiento diferencial, pero que se necesitan tamaños muestrales grandes para hacerlo. Además, resultan más precisos en la detección de diferencias en las cargas factoriales de los ítems cuando éstas son mixtas, es decir, no siempre el mismo grupo presenta factores de carga más bajos en todos los ítems con pérdida de equivalencia.

French y Finch (2006) manipulan el tamaño muestral, el número de factores, el número de indicadores por factor y la distribución de las variables es un estudio de simulación con datos dicotómicos, encontrando que la prueba estadística $\chi^2$ controla adecuadamente el error de tipo I, con un alto poder de detección cuando realiza una estimación por máxima verosimilitud, pero no cuando realiza una estimación por mínimos

cuadrados ponderados robusto (DWLS) que resulta tener un bajo rendimiento. Sin embargo, Elosua (2011) obtiene buenos resultados utilizando el modelo factorial común para datos ordinales, con el método de estimación de mínimos cuadrados ponderados (WLS) en otro estudio de simulación, en el que manipulan el tamaño muestral, el tipo de DIF, la cantidad de DIF y la presencia de impacto. En consonancia con Cheung y Rensvold (2002) y Coenders, Batista-Foguet y Saris (2005) aconseja añadir a la comparación de $\chi^2$ la diferencia del índice CFI entre los modelos, encontrando en su estudio que reduce los falsos positivos.

## 5. EQUIVALENCIA DE MEDIDA DE UNA PRUEBA CON PROCEDIMIENTOS TRI

La TRI proporciona un atractivo marco de trabajo para estudiar la equivalencia de medida, debido a sus propiedades de invarianza: establecen modelos en los que las mediciones obtenidas no varíen en función del instrumento utilizado (invariantes respecto al test) ni en función de los objetos medidos (invariantes respecto de los sujetos). Así, la TRI puede señalar diferencias en el funcionamiento de ítems y de tests sin que este resultado esté afectado por las diferencias en la distribución del rasgo en los grupos que están siendo comparados (Embretson y Reise, 2000). Cuando se encuentra que la relación ítem-rasgo es diferente entre los grupos, entonces el ítem en cuestión presenta funcionamiento diferencial (DIF), y cuando la relación test-rasgo es diferente, el test presenta funcionamiento diferencial (DTF).

El marco de trabajo de la TRI propone un modelo logístico para describir las relaciones entre las respuestas observadas a los ítems y el nivel del rasgo latente, $\theta$. La naturaleza exacta de este modelo se determina por un conjunto de parámetros de los ítems que son potencialmente únicos para cada ítem. Así, esta teoría relaciona características de los ítems (parámetros) y características de los individuos (rasgos latentes) con la probabilidad de elegir cada una de las categorías de respuesta. Esta relación probabilística se define matemáticamente en términos de la función de respuesta al ítem, que es una regresión no lineal de la probabilidad de elegir una categoría de respuesta de un ítem para un nivel de rasgo determinado ($\theta$). Hay varias familias de funciones de respuesta al ítem para modelos unidimensionales o multidimensionales y con formatos de respuesta dicotómico o politómico (para una revisión ver, por ejemplo, De Ayala, 2009; Hamilton y Swaminathan, 1985; Lord, 1980; Embretson y Reise, 2000).

La primera generación de modelos TRI se desarrolló para ser aplicada a tests unidimensionales de rendimiento, habilidades y aptitudes, cuyos ítems estaban puntuados de forma dicotómica (p.e. Birnbaum, 1968; Lord, 1952; Rasch, 1960). Las funciones de respuesta al ítem incorporaban uno, dos o tres parámetros y estaban basadas en dos formas matemáticas, la ojiva normal y la logística.

El sistema de asignación de puntuaciones dicotómico es muy restrictivo y no parece adecuado para recoger toda la información disponible en la mayoría de las aplicaciones (e.g. Donoghue, 1994), motivo por el cual, en la actualidad, la mayoría de los tests de personalidad utilizan para su respuesta una escalas con ítems politómicos, tal y como se ha hecho en la parte empírica de este trabajo.

Antes de proceder al estudio de la equivalencia de medida de un test en varios grupos hay que seleccionar el modelo de TRI a utilizar. En tests que contienen ítems politómicos de respuesta graduada, conceptualmente el modelo TRI más apropiado es un modelo para categorías de respuestas ordenadas. Entre los modelos de este tipo se incluyen el modelo de respuesta graduada de Samejima (1969), el modelo de crédito parcial de Thissen y Steinberg (1986) y el modelo de crédito parcial generalizado de Muraki (1996).

Algunos autores sugieren que estos modelos trabajan de forma similar en situaciones prácticas dado que tienen formas similares (Maydeu-Olivares, Drasgow y Mead, 1994) y otros abogan por la utilización del modelo de respuesta graduada (Bolt, Hare, Vitale y Newman, 2004; Samejima, 1997), siendo este último probablemente el más ampliamente utilizado (Stark, Chernyshenko, Lancaster, Drasgow y Fitzgerald, 2002).

Para una revisión los modelos politómicos de la TRI puede consultarse el libro de Van der Linden y Hambleton (1997), o un volumen especial del Applied Psychological Measurement, editado por Drasgow (1995a); y en castellano, el apartado 4.4.1. del libro Modelos Psicométricos de Santisteban y Alvarado (2001), el capítulo 7 del manual de Psicometría de Martínez, Hernández y Hernández (2006) o el libro monográfico de Revuelta, Abad y Ponsoda (2006).

## 5.1. MODELO DE RESPUESTA GRADUADA DE SAMEJIMA

El Modelo de Respuesta Graduada de Samejima (MRG) se considera una generalización del modelo logístico de dos parámetros de Birnbaum (1968). La estrategia

que permitió a Samejima (1969) la aplicación de este modelo a ítems politómicos consiste en dividir la variable de respuesta politómica en una serie de variables dicotómicas y en especificar una función característica para cada una de ellas. Así, en el MRG, la relación entre la probabilidad de una persona con un nivel de rasgo latente $\theta$ de elegir una opción de respuesta particular a un ítem puede describirse gráficamente con la Curva de la Categoría de Respuesta (CCR) cuya función es:

$$P_{ik}(\theta_s) = \frac{e^{a_i(\theta_s - b_{ik+1})} - e^{a_i(\theta_s - b_{ik})}}{\left(1 + e^{a_i(\theta_s - b_{ik})}\right)\left(1 + e^{a_i(\theta_s - b_{ik-1})}\right)} \qquad (16)$$

donde:

$P_{ik}$ ($\theta$) representa la probabilidad de que un examinado ($s$) con un nivel dado en el rasgo latente ($\theta$) responda al ítem $i$ con la categoría $k$

$a_i$ es la pendiente o parámetro de discriminación

$b_i$ es el parámetro de umbral entre categorías (habrá k-1 parámetros de localización)

$k$ es el número de opciones del ítem $i$.

Un ejemplo de un grafico con las CCR a un ítem de 4 categorías puede verse en la Figura 2 en la que cada línea corresponde a la probabilidad de responder a una de las cuatro categorías de respuesta para el ítem en función del nivel de $\theta$. En el grafico, la línea más a la izquierda corresponde a la probabilidad de responder a la opción de respuesta que indica menor nivel de rasgo; la función de respuesta de esta primera categoría es monótona decreciente, la función correspondiente a la última categoría es monótona creciente (mayor probabilidad cuanto mayor es el nivel de rasgo) y las de las categorías centrales son unimodales (serán los sujetos con un cierto nivel central de rasgo los que más probabilidad tienen de seleccionarlas).

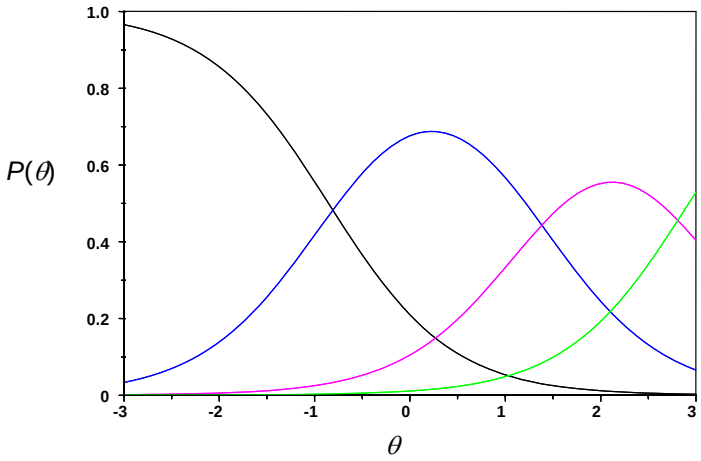



Figura 2. CCR de un ítem tipo Likert con cuatro alternativas.

Samejima utilizó un procedimiento acumulativo, en el que la curva característica de la categoría "k" indica la probabilidad de alcanzar esa categoría o las siguientes, condicionada a la localización del sujeto en el rasgo ($P(X_i \geq k \mid \theta)$). Así, en el MRG, la relación entre el nivel de rasgo latente de un sujeto ($\theta$) y la probabilidad del sujeto de elegir progresivamente un incremento en la categoría de respuesta observada puede representarse por una serie de Curvas Características Operantes (CCO).

La formulación de esta función de probabilidad acumulada es:

$$P_{ik}^{*}(\theta_s) = \frac{e^{a_i(\theta_s - b_{ik})}}{1 + e^{a_i(\theta_s - b_{ik})}} \qquad (17)$$

donde:

$P_{ik}^{*}(\theta_s)$ representa la probabilidad de que un examinado $s$ con un nivel en el rasgo $\theta$ responda en el ítem $i$ a la categoría $k$ o a una categoría por encima de ella.

Cada una de las CCO representa, por tanto, la probabilidad de elegir una categoría igual o superior a k, que se incrementa con el nivel de rasgo. El parámetro *a* de discriminación del ítem estará relacionado con la pendiente en $\theta = b$, en la expresión anterior. Los parámetros de localización determinarán la separación entre las curvas de la Figura 3; un valor de *b* concreto indica el valor del nivel de rasgo para el que es .5 la probabilidad de elegir la alternativa *k* o alguna superior. Cada ítem solo tiene un parámetro *a*, porque éste tiene que ser igual en todas las funciones dentro de un ítem, aunque puede variar de un ítem a otro. En cuanto al parámetro *b*, hay uno menos que categorías de respuesta tiene el ítem, por tanto uno para cada CCO.

Para un ítem con $k_i$ categorías de respuesta , las $(k_i - 1)$ funciones de respuesta límite pueden expresarse como:

$$P_{i1}^{*} = \frac{e^{a_i(\theta_s - b_{i1})}}{1 + e^{a_i(\theta_s - b_{i1})}} \qquad (18)$$

.
.
.

$$P_{i(k_i-1)}^{*} = \frac{e^{a_i\left(\theta_s - b_{i(k_i-1)}\right)}}{1 + e^{a_i\left(\theta_s - b_{i(k_i-1)}\right)}} \qquad (19)$$

Como puede observase en la ecuación 17, cada función de probabilidad acumulada representa un modelo TRI logístico de dos parámetros en el caso de datos dicotómicos. Si $a = 1$, entonces la CCR se convierte en un modelo de Rasch (Hambleton, Swaminathan y Rogers, 1991). Un ejemplo de CCO para un ítem de 4 categorías de respuesta se muestra en la Figura 3.

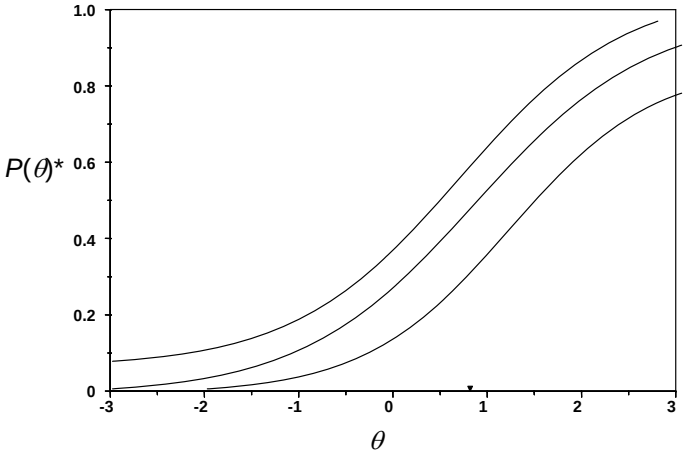



Figura 3. CCO para un ítem de escala Likert con cuatro categorías de respuesta.

Este procedimiento permite estimar la probabilidad condicionada de que un sujeto seleccione una categoría concreta $k$ a partir de la diferencia ($P(X_i \geq k \mid \theta)$)- ($P(X_i \geq k+1 \mid \theta)$), motivo por el cual Thissen y Steinberg (1986), autores de una de las clasificaciones más conocidas de los modelos de la TRI, denominan a los modelos que emplean esta estrategia "modelos diferenciales". Otra clasificación importante es la realizada por Mellenberg (1995), quien agrupa estos modelos bajo la denominación de "respuesta acumulativa".

Las CCR (ver ecuación 16) se derivan de las ecuaciones 18 y 19 y pueden expresarse en función de las diferencias de las probabilidades de dos categorías adyacentes, de la siguiente manera:

$$P_{i1} = 1 - P_{i1}^{*} \qquad (20)$$

$$P_{i2} = P_{i1}^{*} - P_{i2}^{*} \qquad (21)$$

$$\vdots$$

$$P_{i(k_i-1)} = P_{i(k_i-2)}^{*} - P_{i(k_i-1)}^{*} \qquad (22)$$

$$P_{i(k_i)} = P_{i(k_i-1)}^{*} \qquad (23)$$

Hay que tener en cuenta que las CCO y las CCR de un ítem dado dependen de $\theta$ y de los parámetros de los ítems (*a* y *b*).

Utilizando esta información, la puntuación directa esperada o puntuación verdadera (según nomenclatura de Raju *et al*., 2002) en un ítem $i(t_i)$ de un examinado con un determinado nivel de $\theta$ puede expresarse como:

$$t_i(\theta) = (1)P_{i1} + (2)P_{i2} + ... + (k_i - 1)P_{i(k_i-1)} + (k_i)P_{i(k_i)} \quad (24)$$

que, dadas las ecuaciones 20-23, se puede simplificar a la siguiente expresión:

$$t_i(\theta) = 1 + P_{i1}^* + ... + P_{i(k_i-1)}^* \quad (25)$$

siendo en ambas ecuaciones las categorías definidas como 1, 2, ..., $k_i - 1$, y $k_i$.

La puntuación verdadera o esperada (T) de la subescala o test que contiene *n* ítems politómicos puede expresarse como:

$$T(\theta) = \sum_{i=1}^{n} t_i(\theta) \quad (26)$$

Dadas las ecuaciones 25 y 26, las puntuaciones observadas del ítem y subescala o test pueden expresarse como:

$$x_1 = t_1(\theta) + e_i \quad (27)$$

.
.

.

$$x_n = t_n(\theta) + e_n \tag{28}$$

y

$$x = (x_1 + \ldots x_n) = \sum_{i=1}^{n} t_i(\theta) + \sum_{i=1}^{n} e_i = T(\theta) + e \tag{29}$$

donde *e* representa el componente error.

En vista de las ecuaciones 27-29, la esperanza y la varianza de la puntuación observada del ítem es:

$$E(x_i) = t_i(\theta) \tag{30}$$

$$\sigma^2_{x_i} = \sigma^2_{t_i} + \sigma^2_{e_i} \tag{31}$$

Siendo a nivel de subescala o test completo:

$$E(x) = T(\theta) \tag{32}$$

$$\sigma^2_x = \sigma^2_T + \sigma^2_e \tag{33}$$

Se observa que las ecuaciones 27-33 son muy similares en formato a las ecuaciones 6-10. Por ejemplo, las ecuaciones 27 y 28 muestran que una puntuación observada de un ítem es simplemente la suma de la puntuación directa esperada (o puntuación verdadera) y la puntuación error (modelo básico de la TCT) y la ecuación 5 se define de manera similar en el contexto del AFC. Además, las ecuaciones 31 y 33 en el modelo TRI y las ecuaciones 8 y 10 en el modelo de AFC reflejan el supuesto de que las puntuaciones verdaderas y error no están correlacionadas, que es un supuesto subyacente a la TCT. La mayor diferencia entre estos dos pares de ecuaciones es la forma en la que se ha definido la

puntuación verdadera, (para profundizar sobre el grado de similitud entre las estructuras subyacentes de TRI, AFC y TCT, ver McDonald, 1999). En consecuencia, existe una relación entre AF y TRI que debería reflejarse en los resultados que se obtienen en una y otra metodología en el análisis de la equivalencia de medida.

Las ecuaciones 30 y 32 hacen referencia a la función de respuesta al ítem (o puntuación esperada en el ítem) y a la Curva Característica del Test (CCT) (o puntuación esperada en el test), respectivamente.

## VALORACIÓN DEL AJUSTE DEL MODELO

La evaluación de la invarianza de medida mediante un procedimiento basado en la TRI se enfrenta a un problema importante: ¿cómo evaluar el ajuste del modelo a los datos? Y es que la TRI, a diferencia de otros modelos, no finaliza con el proceso de estimación de parámetros, sino que requiere la realización de un estudio de bondad de ajuste del modelo matemático especificado. En este sentido, se dice que la TRI se basa en supuestos matemáticos fuertes, ya que requiere y permite una comprobación empírica de su cumplimiento.

Un problema fundamental en la evaluación de los modelos es la sensibilidad de los estadísticos de bondad de ajuste al tamaño muestral del estudio, que hace posible que, incluso pequeños desajustes resulten estadísticamente significativos en muestras grandes, llevando al rechazo de la hipótesis nula. Por el contrario, si se eligen tamaños muestrales pequeños, los estimadores de los parámetros resultan de escasa calidad por su inconsistencia y el gran tamaño de los errores típicos, ya que el uso de máxima

verosimilitud como procedimiento de estimación de parámetros tiene como condición que n→∞ (Santisteban, 1990). De hecho, la evaluación de la bondad de ajuste en los modelos de respuesta al ítem es una cuestión que continúa suscitando debate en la literatura (Chernyshenko, Stark, Chan, Drasgow y Williams, 2001), siendo considerablemente más difícil que en los modelos de AFC (McDonald y Mok, 1995; Reise, Widaman y Pugh, 1993). Puede consultarse una revisión de los procedimientos de valoración del ajuste de un modelo TRI a los datos en Hambleton (1989), en Hambleton y Swaminathan (1985), en López-Pina e Hidalgo (1996) o en Swaminathan, Hambleton y Rogers (2007).

Drasgow, Levine, Tsien, Williams y Mead (1995) utilizan una combinación de métodos gráficos y estadísticos complementarios para evaluar el ajuste de los modelos, que ha sido bien acogida en la literatura (Bolt *et al.*, 2004; Chernyshenko *et al.*, 2001; Robie, Zickar y Schmit, 2001; Stark *et al.*, 2002).

Análisis gráfico

La idea del ajuste gráfico es representar las funciones de respuesta a un ítem, estimadas en una muestra de calibración, así como las proporciones empíricas de respuestas obtenidas en una muestra de validación. En la versión más simple, un ajuste gráfico se construye dividiendo el continuo $\theta$ en, por ejemplo, 25 estratos. Después se estima $\theta$ para cada sujeto, y se cuenta el número total de sujetos en cada estrato. Una proporción empírica se computa como el número de sujetos que seleccionan la opción dividido por el número total de sujetos del estrato. El procedimiento de simple suma de Samejima (1983) proporciona un ejemplo tradicional de ajuste gráfico. Así, $P_i(t)$ de una función de respuesta al ítem se computa como:

$$\hat{P}_i(t) = \frac{\sum_{(A:A\in S^+)} P(\theta = t|\hat{\tau} = \hat{\tau}_A)}{\sum_{A} P(\theta = t|\hat{\tau} = \hat{\tau}_A)} \qquad (34)$$

donde el sumatorio en el denominador abarca a todos los sujetos de la muestra, el sumatorio del numerador únicamente a los sujetos que responden correctamente al ítem i, $\hat{\tau}$ es un estimador de $\theta$ que se calcula sobre todas las respuestas excepto el ítem en cuestión, y $\hat{\tau}_A$ es el valor de $\hat{\tau}$ calculado con los datos del sujeto $A$.

El problema de este procedimiento de suma directa es que la estimación de $\theta$ $(\hat{\tau})$ para el sujeto casi nunca es igual a la real debido a un error de estimación. El error en $\hat{\tau}$ puede cambiar el ajuste gráfico de tal manera que incluso con una muestra muy grande y funciones de respuesta perfectamente estimadas, el ajuste gráfico puede diferir sustancial y sistemáticamente de la función de respuesta verdadera. Este problema es especialmente pronunciado en test cortos donde la estimación de $\theta$ puede conllevar más error.

En el modelo de Samejima, la simple suma estimada de un punto sobre la función de respuesta al ítem/opción, $P_i(t)$ se computa con la estimación de $\hat{\tau}$. Una solución a este problema es remplazar esta estimación con el estadístico del vector, que es simplemente el patrón de respuesta de los sujetos $u^*$ a un conjunto dado de ítems (incluido el ítem objetivo). La estimación empírica de una función de respuesta al ítem puede ser escrita como:

$$\hat{P}_i(u_i = 1|\theta = t) = \frac{N^+}{N} \frac{\sum_{(A:A\in S^+)} P(\theta = t|u = u_A^*)/N^+}{\sum_{A} P(\theta = t|u = u_A^*)/N} \qquad (35)$$

donde:

$N^+$ es el número de sujetos que han respondido al ítem correctamente (o a una categoría determinada); N es el número total de sujetos; $u^*_A$ es el patrón de respuestas dicotomizado del sujeto A; y $S^+$ es el conjunto de sujetos que han respondido al ítem correctamente.

Análisis estadístico

Las pruebas estadísticas de bondad de ajuste (por ejemplo el estadístico de ajuste $\chi^2$) son probablemente las más utilizadas en la evaluación del ajuste del modelo a los datos. Desafortunadamente, su sensibilidad al tamaño muestral y su insensibilidad a ciertas formas de desajuste del modelo a los datos no suele conducir a conclusiones inequívocas acerca del ajuste del modelo a los datos. Un método mejorado de computar este estadístico son los estadísticos $\chi^2$ ajustados al ratio de los grados de libertad. Estos índices evalúan el ajuste del MRG de Samejima respecto a las frecuencias conjuntas de primer orden, segundo orden y tercer orden respectivamente de las puntuaciones de los ítems. Estos estadísticos $\chi^2$ son índices que cuantifican la diferencia entre el número esperado de respuestas a una opción del ítem (derivado de la CCR teórica) y las frecuencias observadas de las respuestas dadas a esa opción en el conjunto de ítems.

Los estadísticos $\chi^2$ se computan para cada ítem individualmente y para conjuntos de dos y tres ítems. Hay *n* estadísticos $\chi^2$ que se calculan para los *n* ítems individuales de la prueba. Sin embargo, hay $\binom{n}{2}$ estadísticos $\chi^2$ que se calculan para los conjuntos de dos ítems y $\binom{n}{3}$ estadísticos $\chi^2$ para cada triplete de ítems. Estos conjuntos de ítems se seleccionan de tal manera que cada uno contenga un ítem relativamente fácil, un ítem de moderada dificultad y un ítem relativamente difícil.

El $\chi^2$ para un ítem *i* se calcula teniendo en cuenta las frecuencias observadas y esperadas:

$$\chi_i^2 = \sum_{k=1}^{s} \frac{[O_i(k) - E_i(k)]^2}{E_i(k)} \qquad (36)$$

donde *s* es el número de opciones, $O_i(k)$ es la frecuencia observada de la opción *k*, y $E_i(k)$ es la frecuencia esperada de la opción k bajo el modelo TRI especificado. La frecuencia esperada de elegir una opción se calcula con la siguiente fórmula:

$$E_i(k) = N \int P(v_i = k|\theta = t) f(t) dt \qquad (37)$$

Donde N es el número de sujetos y f(·) es la densidad θ, habitualmente tomada de la normal estandarizada porque las funciones de respuesta a la opción (o al ítem) se escalan en referencia a esta distribución. Esta integral se evalúa mediante cuadratura numérica, utilizando 61 puntos de anclaje en el intervalo (-3, +3). Para pasar por alto la sensibilidad al tamaño de la muestra y para permitir comparaciones entre distintas muestras y tests, se ajusta $\chi^2$ a la magnitud que se esperaría con una muestra de 3000 personas. Entonces se calcula la razón del estadístico $\chi^2$ entre los grados de libertad. Un valor de más de 3 para cualquier ítem indica un desajuste del modelo a los datos.

Como se ha visto, los ítems simples se computan basándose en el número esperado de veces que los sujetos seleccionarían la opción *k* dadas las probabilidades del modelo TRI. Los estadísticos $\chi^2$ para conjuntos de dos ítems se computan basándose en las probabilidades esperadas y observadas de presentar opciones específicas de respuesta en dos ítems (la tabla de contingencia compara las probabilidades esperadas y observadas de elegir la opción 1 en el ítem 1 y la opción 2 en el ítem 2, etc.). Los ítems triples se computan de forma similar con una tabla de contingencia de tres vías.

Los estadísticos de $\chi^2$ para ítems individuales son, en muchas ocasiones, insensibles a ciertos tipos de desajuste, como la violación del supuesto de unidimensionalidad. Para evitar este problema, su cálculo se complementa con los conjuntos de dos y tres ítems, cuyo estadístico $\chi^2$ capta este tipo de desajuste. Además, calculando los índices de $\chi^2$ para conjuntos de dos y tres ítems se puede valorar la capacidad del modelo TRI para predecir la interacción entre los ítems. Para escalas de las que se sospecha multidimensionalidad el examen de estas interacciones entre ítems es necesario.

## 5.2. EQUIVALENCIA DE MEDIDA EN DIVERSOS GRUPOS EN EL ÁMBITO DE LA TRI

En la TRI, realizar pruebas de equivalencia de medida supone, esencialmente, determinar si los parámetros *a* y *b* son equivalentes en los grupos, utilizando los métodos disponibles para la evaluación del funcionamiento diferencial de ítems y de tests (Raju, *et al*., 1995).

Se dice que un ítem presenta equivalencia de medida si los parámetros del ítem permanecen invariantes en las dos poblaciones. Esto es, a un nivel de ítem:

$$a_i^g = a_i^{g'},\ \ b_{i1}^g = b_{i1}^{g'},\ ...,\ b_{ik_i}^g = b_{ik_i}^{g'} \qquad (38)$$

donde g´ representa la segunda población.

Cuando los parámetros del ítem son iguales, las CCR y las CCO de un ítem también son iguales para las dos poblaciones. Además, las puntuaciones verdaderas del ítem (ver

ecuación 30) son iguales para dos personas con idénticas puntuaciones en la variable latente $\theta$.

Estudiando los efectos acumulativos del DIF en los diferentes ítems del test se puede indagar si las puntuaciones del test representan niveles diferentes del rasgo estudiado entre los grupos, lo que implicaría una pérdida de equivalencia en la escala o subescala completa.

Es importante destacar que la existencia de ítems que presentan funcionamiento diferencial no implica necesariamente una pérdida de equivalencia en la escala. Esta afirmación se basa en estudios como el de Drasgow (1987) ya comentado anteriormente, y más recientemente por Cooke, Kosson y Michie (2001), que utilizando un test para evaluar psicopatía (Psychopathy Checklist-Revised) comparan delincuentes afroamericanos y caucásicos, encontrando que, de los 20 ítems del test, 5 presentan DIF. Sin embargo, el funcionamiento diferencial observado en los ítems ocurre en direcciones opuestas por lo que su efecto se anula al nivel de puntuación total en el test y los autores sugieren que el efecto global del DIF en las puntuaciones totales es insignificante.

Hay varios procedimientos DIF basados en la TRI: el $\chi^2$ de Lord (1980), las medidas del área de Raju (1988, 1990), el test de razón de verosimilitud de Thissen *et al*., (1988), y los procedimientos de Raju *et al.* (1995) basados en el funcionamiento diferencial de ítems y tests (DFIT).

El $\chi^2$ de Lord y las medidas del área de Raju son procedimientos que se proponen inicialmente para la evaluación del funcionamiento diferencial del ítem dentro de los

modelos TRI dicotómicos, y que, posteriormente Cohen, Kim y Baker (1993) amplían para incluir los modelos TRI politómicos. Los procedimientos de Thissen *et al.* (1993) y de Raju *et al.* (1995) son apropiados para la evaluación de DIF tanto con puntuaciones dicotómicas como politómicas (Flowers *et al.*, 1999). Además, el procedimiento de Raju *et al.* (1995) también es apropiado para los modelos TRI multidimensionales (Oshima, Raju y Flowers, 1997).

En cuanto a procedimientos gráficos basados en la TRI, goza de gran aceptación el estudio de las funciones de respuesta esperada (Bolt *et al*. 2004). Una función de respuesta esperada representa la puntuación del ítem esperada como una función de θ y se calcula como la suma de las categorías de puntuación del ítem ponderado por sus probabilidades (ver ecuación 24). Comparar las funciones de respuesta esperada en varios grupos es una manera útil de interpretar el funcionamiento diferencial del ítem en varias poblaciones. Cuando unos datos se ajustan a un modelo de la TRI, existe DIF si los parámetros de un ítem tienen diferentes valores en los distintos grupos y, en ese caso, las funciones de respuesta esperada serán diferentes necesariamente (Chang y Mazzeo, 1994).

Las funciones de respuesta esperada son más fáciles de analizar visualmente que las curvas características del ítem, ya que solo hay una curva por ítem. Por este motivo, se consideran una atractiva forma de valorar las implicaciones del funcionamiento diferencial en las puntuaciones esperadas de interés (Bolt *et al*. 2004). Además, pueden proporcionar las bases para cuantificar la cantidad de DIF existente en un ítem. Por ejemplo, Cohen, Kim y Baker (1993) y Wainer (1993) comentan índices basados en el signo y la distancia entre las funciones de respuesta esperada de dos grupos como una forma de cuantificar el funcionamiento diferencial.

Aunque los procedimientos basados en la TRI son técnicas muy utilizadas e importantes en el estudio del funcionamiento diferencial (Budgell, Raju y Quartetti, 1995; Cohen *et al*, 1993; Drasgow y Hulin, 1990; Millsap y Everson, 1993; Raju, 1988, 1990; Raju *et al.*, 1995), tradicionalmente se han limitado a medir el funcionamiento diferencial a nivel de ítem.

El funcionamiento diferencial a nivel de test, sin embargo, ha merecido una menor atención (Collins, Raju y Edwards, 2000), no habiendo apenas índices que lo midan. Una excepción es el procedimiento DFIT, desarrollado por Raju *et al.* (1995) que contiene índices que evalúan el DIF y un índice para evaluar el DTF. Otro procedimiento TRI que permite comprobar la equivalencia de medida a nivel de test es la comparación de modelos basada en el test de razón de verosimilitud de Thissen *et al*., (1988).

## 5.3. COMPARACIÓN DE MODELOS BASADA EN LA RAZÓN DE VEROSIMILITUDES

La comparación de modelos basada en la razón de verosimilitudes (Likelihood Ratio; LR) implica la comparación del ajuste de dos modelos: un modelo compacto con restricciones que establece la igualdad de los parámetros de los ítems, con un modelo base o aumentado en el que se asume que los parámetros de los ítems del test pueden diferir entre los grupos (Thissen *et al*, 1986; Thissen *et al*, 1988; 1993). Por tanto, se dispone de una hipótesis nula, que plantea que los datos se ajustan al modelo compacto, y de una hipótesis alternativa, que plantea que los datos se ajustan al modelo aumentado. El objetivo

es probar si el modelo aumentado mejora significativamente el ajuste de los datos y el estadístico $G^2$ utilizado para comparar los modelos es el logaritmo neperiano de una razón de verosimilitudes dada por:

$$G^2 = -2\ln\frac{L_C}{L_A} \qquad (39)$$

donde:

$L_c$ es la función de verosimilitud del modelo compacto (que contiene menos parámetros)

$L_A$ es la función de verosimilitud del modelo aumentado, en el que se permite que los parámetros de los ítems varíen de un grupo a otro.

Este estadístico de contraste sigue una distribución $\chi^2$ con grados de libertad igual a la diferencia en el número de parámetros entre el modelo aumentado y el modelo compacto (Hidalgo y Gómez, 1999; Teresi *et al*., 2007). Si el valor obtenido es menor que el valor teórico de la distribución, no hay evidencias de diferencias en el ajuste de ambos, lo que apoyaría la equivalencia de medida de la prueba en la variable estudiada. Si, por el contrario, el valor del estadístico es mayor que el valor teórico de la distribución hay diferencias en el ajuste de ambos modelos por lo que no hay invarianza o equivalencia; será el momento de buscar qué ítems son los causantes del desajuste en el marco de la equivalencia parcial de medida.

Aunque como indican algunos autores (Thissen *et al*., 1986; Wainer, Sireci y Thissen, 1991) esta estrategia puede ser utilizada, tal y como se acaba de presentar, para comprobar la equivalencia de medida en un test completo, lo cierto, es que surgió y se ha utilizado mayoritariamente para comprobar el DIF (Cohen, Kim y Baker, 1993; Cohen,

Kim y Wollack, 1996; Thissen, 1991; Thissen, *et al*., 1988; 1993). En los trabajos de Haberman (1977) ya se sugiere la utilización del estadístico $G^2$ para evaluar el ajuste de un modelo invariante de medida entre grupos respecto a un modelo no invariante.

Para evaluar el DIF se utiliza en este caso el mismo procedimiento de comparación de modelos, con un modelo compacto que establece la igualdad de parámetros en todos los ítems excepto en el ítem objeto de estudio. Por tanto, la aplicación para el estudio del DIF es similar. Primero, el modelo base se evalúa en todos los parámetros de los ítems y para todos los ítems del test con una única restricción: la igualdad de los parámetros de los ítems en ambos grupos, es decir entre el ítem 1 del grupo 1 y del grupo 2. Este modelo compacto proporciona un valor de verosimilitud base para el ajuste de los parámetros de los ítems al modelo.

Después, para evaluar el DIF de cada ítem, se ejecuta el análisis una vez para cada uno de ellos, con la restricción de que todos los parámetros de los ítems tienen que ser iguales en los grupos, con excepción de los parámetros del ítem del que se evalúa su funcionamiento diferencial. Este modelo aumentado proporciona un valor de verosimilitud asociado con la estimación de los parámetros para el ítem *i* por separado para cada grupo.

Esta prueba de razón de verosimilitud puede realizarse con el programa MULTILOG (Thissen, 1991) aunque su cálculo es muy laborioso y requiere múltiples ejecuciones del programa. Thissen (2001) ha implementado de manera más manejable esta prueba en su programa IRTLRDIF.

Este procedimiento de detección de funcionamiento diferencial ha resultado ser bastante eficaz. Por ejemplo, Cohen, Kim y Wollack (1996) han examinado la calidad del test LR para detectar funcionamiento diferencial del ítem bajo una variedad de situaciones utilizando datos simulados, concluyendo que el índice se comportaba razonablemente bien. Algunos autores como Meade (2010) consideran, sin embargo, que este procedimiento tiene la desventaja de detectar incluso diferencias muy pequeñas en el funcionamiento del ítem cuando los tamaños muestrales son grandes.

Además, en lo que respecta al estudio empírico que se presenta en esta investigación, tiene la ventaja de que el estadístico $G^2$ facilita la comparación entre los procedimientos basados en AFC y TRI (Hambleton, Swaminathan y Rogers, 1991; Reise, Widaman y Pugh, 1993; Scandura, Williams y Hamilton, 2001), por la similitud de la forma de trabajar de ambos procedimientos (ambos comparan un modelo base con un modelo con restricciones).

## 5.4. PROCEDIMIENTO BASADO EN EL FUNCIONAMIENTO DIFERENCIAL DE ÍTEMS Y TESTS (DFIT)

Este marco de trabajo, desarrollado por Raju, Van der Linden y Fleer (1995) proporciona medidas basadas en la TRI del funcionamiento diferencial a nivel de ítems y de tests (o subescalas).

Raju *et al.* (1995) utilizan el término puntuación verdadera del ítem que, en la TRI, es simplemente la puntuación directa esperada en función de la probabilidad de obtener la

respuesta correcta. Dados los parámetros del ítem para el grupo focal y grupo de referencia, se pueden computar dos puntuaciones verdaderas para cada persona: una puntuación verdadera utilizando los parámetros del ítem del grupo focal y otra utilizando los del grupo de referencia. Estas dos puntuaciones son idénticas cuando los parámetros del grupo focal y del grupo de referencia son iguales, esto es:

$$d_{is} = t_{is_F} - t_{is_R} = 0 \qquad (40)$$

para todos los valores de $\theta$, siendo:

$d_{is}$ = diferencia entre las puntuaciones verdaderas en el ítem *i* del sujeto *s*, considerando que pertenece al grupo focal y al grupo de referencia

$t_{isF}$ = puntuación verdadera en el ítem *i* del sujeto *s*, considerado del grupo focal

$t_{isR}$ = puntuación verdadera en el ítem *i* del sujeto *s*, considerado del grupo de referencia.

Esto significa que las funciones de respuesta al ítem en el grupo focal y de referencia son idénticas para el ítem *i*. De forma similar, la diferencia en puntuación verdadera a nivel de test para una persona *s* puede definirse como:

$$D_s = \left(T_{s_F} - T_{s_R}\right) = d_{1s} + \ldots + d_{ns} \qquad (41)$$

donde:

$$T_{s_F} = t_{1s_F} + \ldots + t_{ns_F} \qquad (42)$$

$$T_{s_R} = t_{1s_R} + \ldots + t_{ns_R} \qquad (43)$$

La ecuación 42 representa la puntuación verdadera en el test para una persona *s* del grupo focal, y la ecuación 43 representa la puntuación verdadera en el test de la misma persona si fuera del grupo de referencia. En cada caso, la puntuación verdadera en el test es

simplemente la suma de las puntuaciones verdaderas en cada uno de los *n* ítems del test. Las ecuaciones 42 y 43 se refieren a las funciones de respuesta del test.

La equivalencia de medida a nivel de puntuación total en la escala o subescala implica que $D_s = 0$ para todos los valores de θ o para todas las personas. Además, la equivalencia de medida en DFIT significa que las diferencias en puntuación verdadera son iguales a cero a nivel de ítem y de subescala. La equivalencia de medida está siempre garantizada cuando los parámetros del ítem son iguales en las dos subpoblaciones. En la práctica, la evaluación de la equivalencia de medida utilizando el procedimiento DFIT gira en torno al grado en que *d* y *D* son significativamente distintos de cero.

Este procedimiento incluye una medida del funcionamiento diferencial del test (Differential Test Functioning; DTF) y dos medidas del DIF, denominadas funcionamiento diferencial compensatorio del ítem (Compensatory Differential Item Functioning; CDIF) y no compensatorio (Noncompensatory Differential Item Functioning; NCDIF). La asociación entre el funcionamiento diferencial del test y el funcionamiento diferencial compensatorio del ítem es aditiva. Esto es, DTF es la suma de CDIF para todos los ítems del instrumento de medida. Dado que el CDIF de cada ítem se suma para obtener el total DTF, el funcionamiento diferencial de cada ítem es compensatorio, de ahí el nombre CDIF. Además, si un ítem influye a favor del grupo 1 y otro ítem influye de igual forma, pero a favor del grupo 2, el CDIF sumado de estos dos ítems se cancelará uno con otro cuando se combinen para formar el DTF del test total. CDIF, por tanto, también tiene en cuenta el DIF de otros ítems en un instrumento de medida o test.

Utilizando la ecuación 41, Raju *et al*. (1995) definen el índice de funcionamiento diferencial del test (DTF):

$$DTF = E\left(D^2\right) = \mu_{D^2} = \sigma_D^2 + \mu_D^2 \qquad (44)$$

De manera similar, y basándose en la ecuación 40 definen el índice de funcionamiento diferencial del ítem no compensatorio (NCDIF):

$$NCDIF = E\left(d^2\right) = \mu_{d^2} = \sigma_d^2 + \mu_d^2 \qquad (45)$$

NCDIF es una prueba del funcionamiento diferencial a nivel del ítem, que determina si cada ítem funciona diferencialmente en los grupos, independientemente de otros ítems de la escala. NCDIF es un caso especial de CDIF en el que se asume el supuesto de que todos los ítems a excepción del que está siendo estudiado están libres de DIF.

De acuerdo con la ecuación 45, el índice NCDIF refleja la media de la diferencia al cuadrado entre las puntuaciones verdaderas a nivel de ítem, del grupo focal y grupo de referencia. De manera similar y en consonancia con la ecuación 44, el índice DTF es la media de la diferencia al cuadrado en puntuaciones verdaderas a nivel de escala o subescala.

Para calcular estos índices puede utilizarse el programa DFITPUA (Raju, *et al*., 1995). Para interpretarlos, los autores desarrollaron originalmente tests de significación basados en el estadístico $\chi^2$ para NCDIF y DTF. No se realizan pruebas individuales de CDIF, pero si DTF es significativo, el ítem con un valor de CDIF más alto se elimina de la

escala, realizándose un nuevo análisis de DTF y continuando este procedimiento iterativo hasta que DTF deja de ser significativo. Estos tests de significación han resultado ser muy sensibles al tamaño muestral, de manera que en muestras grandes se tiende a identificar más ítems con DIF de los que hay realmente. Basándose en estudios de simulación, Raju recomendó entonces unos puntos de corte predeterminados de NCDIF > 0'006 para ítems dicotómicos y NCDIF > 0'006 $(k - 1)^2$ para ítems politómicos.

Estos puntos de corte han recibido diversas críticas, porque se ha mostrado en estudios de simulación que los puntos de corte apropiados para determinar si existe DIF dependen de factores como el tamaño muestral y el modelo TRI utilizado (Bolt, 2002), por lo que estos valores no deben generalizarse a todas las situaciones (Oshima y Morris, 2008). En un estudio de simulación, Meade, Lautenschlager y Johnson (2007) concluyen que el problema del procedimiento DFIT es su baja sensibilidad para identificar ítems con DIF y recomiendan que se utilicen valores alternativos de puntos de corte para NCDIF.

Recientemente, Oshima, Raju y Nanda (2006) han desarrollado el método de replicación de los parámetros del ítem (IPR) que proporciona un medio de obtener valores de corte que se adapten a un determinado conjunto de datos de respuesta dicotómica. Este método ha sido recientemente ampliado a formatos de respuesta politómica en el estudio de Raju, Fortmann-Johnson, Kim, Morris, Nering y Oshima (2009).

El método IPR comienza con las estimaciones de los parámetros del ítem para el grupo focal y las varianzas y covarianzas muestrales de esas estimaciones. Basándose en estas estimaciones iniciales, realiza un gran número de replicaciones de los parámetros de los ítems con la restricción de que la esperanza de los parámetros de los ítems recién

generados sea igual a las estimaciones iniciales, con la misma estructura muestral de varianzas-covarianzas.

Dado que se generan a partir de la misma distribución, cualquier diferencia en los conjuntos de estimaciones se debe a errores de muestreo. Estas muestras se utilizan entonces para calcular los estadísticos DIF, obteniéndose una distribución muestral empírica de NCDIF bajo la hipótesis nula de que los grupos focal y de referencia tienen parámetros idénticos. Los valores resultantes de NCDIF se clasifican y el punto de corte se establece en el percentil correspondiente al nivel de alfa deseado (por ejemplo, el percentil 99 para $\alpha = 0{,}01$) (ver Raju *et al*., 2009 para consultar la formulación completa del método IPR).

Cabe señalar que este enfoque no se ajusta a las diferencias en el tamaño de la muestra entre el grupo focal y de referencia, que pueden producir matrices de covarianza diferentes, incluso cuando los parámetros del ítem son idénticos. Por lo tanto, utilizar la matriz de covarianzas del grupo focal para representar los dos grupos puede dar lugar a alguna inexactitud cuando los tamaños de muestra son muy diferentes. Sin embargo, tanto en la investigación de Oshima *et al*., (2006) con datos dicotómicos, como en la de Raju *et al.* (2009) con datos politómicos encontraron resultados precisos utilizando el método IPR, incluso con importantes diferencias de tamaño entre el grupo focal y el grupo de referencia.

Los estudios que han contrastado la eficacia del procedimiento de comparación de modelos basado en el test de razón de verosimilitud con el marco de trabajo DFIT han concluido que este último es menos sensible en la detección de funcionamiento diferencial,

tanto de ítems como a nivel de escala completa (Bolt, 2002; Braddy, Meade y Johnson, 2006; Meade y Lautenschlage, 2004c). Hay que tener en cuenta, sin embargo, que estas investigaciones se realizaron con anterioridad a que Raju *et al.* desarrollaran en el año 2009 el método de replicación de los parámetros del ítem (IPR) que proporciona un medio de obtener valores de corte que se adapten a un determinado conjunto de datos de respuesta politómica.

## 6. RELACIONES ENTRE PROCEDIMIENTOS BASADOS EN AFC Y EN TRI PARA ESTABLECER LA EQUIVALENCIA DE MEDIDA EN UN TEST

Los métodos para evaluar la invarianza basados en el AFC y en la TRI son similares conceptualmente pero distintos en la práctica (Raju *et al*., 2002; Reise, Widaman y Pugh, 1993). En primer lugar, se van a examinar las similitudes y diferencias entre unos y otros para pasar seguidamente a ver si existe información única (o privativa) que proporcione alguno de ellos en relación a la equivalencia de dos muestras de sujetos (Zickar y Robie, 1999). Por último, se ofrece una descripción sobre el estado de la cuestión respecto a la comparación de procedimientos basados en ambas aproximaciones en el estudio de la invarianza de medida en dos muestras.

La semejanza más obvia es que ambas perspectivas examinan la relación entre un constructo subyacente y un conjunto de variables observables (puntuación en el ítem o escala) a los que está ligado teóricamente. En este sentido, ambas aproximaciones examinan el grado en el que las puntuaciones esperadas (o verdaderas según denominación

de Raju *et al*., 2002) a nivel de ítem o escala, de sujetos con el mismo nivel del rasgo latente, son similares en las dos poblaciones. Esta es una similitud importante; conceptualmente es similar a una definición de paralelismo en la TCT. En esta teoría, un requisito para el paralelismo es la igualdad de las puntuaciones verdaderas de dos tests, mientras que aquí se refiere a la igualdad de puntuaciones verdaderas o esperadas en las dos poblaciones cuando la puntuación latente se mantiene constante.

La definición de equivalencia de medida no implica que la distribución de las puntuaciones del rasgo subyacente en las dos poblaciones de interés sea la misma. De hecho, las distribuciones latentes pueden ser, y habitualmente lo son, diferentes (lo que se denomina impacto). La definición de equivalencia de medida simplemente significa que las personas con el mismo nivel del rasgo latente tendrán la misma puntuación esperada a nivel de ítem o test, independientemente de la población a la que pertenecen.

Cuando no hay equivalencia de medida, ambas aproximaciones pueden utilizarse para identificar la extensión y la fuente del problema. En el contexto de la TRI, habitualmente se evalúan los ítems individuales para ver si presentan DIF. Sin embargo, en el contexto del AFC se evalúa el modelo propuesto para su bondad de ajuste a los datos separadamente, antes de buscar el origen de la no equivalencia.

Las funciones de respuesta a los ítems pueden ser una fuente de información útil para evaluar la falta de equivalencia de la medida, tanto en el contexto de la TRI como en el del AFC. Estos gráficos se pueden utilizar para identificar el grado y la localización de la no equivalencia de medida, para un ítem o escala determinada.

En cuanto a las diferencias, la más obvia es que en el AFC la relación entre el constructo latente y sus indicadores es lineal, mientras que en la TRI es no lineal. Aunque esta diferencia entre ambos procedimientos es relevante, McDonald (1999) unifica ambas aproximaciones proporcionando información sobre las estructuras lineales subyacentes a los modelos de CFA y TRI.

En este sentido, Lord (1980) mostró que el análisis factorial para datos dicotómicos es equivalente a la ojiva normal de dos parámetros de la TRI. Así, las relaciones entre los parámetros del análisis factorial confirmatorio λ (carga factorial) e $\tau$ (ordenada en el origen o intercepto) y los parámetros de la TRI *b* (dificultad o localización) y *a* (discriminación o pendiente) vienen dados por (ver Lord, 1980; McDonald, 1999; Ferrando y Lorenzo-Seva, 2005):

$$b_i = \frac{\tau_i}{\lambda_i}$$

y

$$a_i = \frac{\lambda_i}{\sqrt{1 - \lambda_i^2}}$$

La equivalencia entre modelos de AF y TRI se logra cuando para realizar el AF se utiliza la matriz de correlaciones tetracóricas (o policóricas en el caso politómico) en lugar de la matriz de correlaciones de Pearson o la matriz de covarianzas.

En los casos en los que se asigna una puntuación dicotómica a los ítems se considera más apropiado un modelo de regresión logístico para expresar la relación entre un constructo subyacente continuo y una variable observada que un modelo de regresión lineal. Por este motivo, en este caso, puede resultar preferible utilizar procedimientos

basados en la TRI para evaluar la equivalencia de medida (Raju *et al*, 2002). No obstante, si el número de puntuaciones posibles para un ítem se incrementa, o si se utiliza para el análisis la matriz de correlaciones tetracórica, el modelo de regresión lineal que se utiliza en el AFC puede ser igualmente apropiado.

Por otra parte, es importante señalar que la metodología del AFC facilita el manejo de rasgos multidimensionales y varias poblaciones de manera simultánea. Sin embargo, gran parte de la metodología para evaluar la equivalencia de medida del DIF basado en TRI requiere escalas unidimensionales y analiza la invarianza de dos en dos grupos. Eso sí, hay que destacar en este sentido, los avances de Kim, Cohen y Park (1995) para análisis DIF en múltiples grupos y por Oshima *et al.* (1997) para DIF multidimensional.

En cuanto al tratamiento del error en ambas aproximaciones, hay que tener en cuenta que, aunque Jöreskog inicialmente propuso evaluar la invarianza del error de medida (en una forma estricta de equivalencia de medida del AFC, se exige que las varianzas error sean iguales en las distintas poblaciones), el requisito de igualdad de varianzas residuales del ítem es extremadamente riguroso y no muy realista en la gran mayoría de situaciones prácticas (Byrne, 1994, 1998, 2001). En la TRI, no hay demasiada discusión explícita sobre la varianza error a nivel de ítem, porque está condicionada a $\theta$, y en el caso dicotómico, esta varianza puede expresarse como $P_i(\theta)$ (1- $P_i(\theta)$), para el ítem $i$, donde $P_i(\theta)$ representa la probabilidad de responder al ítem $i$ correctamente dado un nivel determinado de $\theta$. El concepto que sí ha recibido mayor atención en TRI es el error estándar de medida asociado a una estimación de $\theta$; el hecho de que varíe como una función de $\theta$ significa que el error estándar de medida puede variar de persona a persona. Esto constituye un gran beneficio de la TRI sobre la TCT (Hambleton, Swaminathan y

Rogers, 1991). Es posible que los promedios de las varianzas residuales de los sujetos de las poblaciones de interés sean iguales, aunque esta relación no ha recibido atención en los estudios de equivalencia basados en la TRI. La igualdad de estos promedios de varianzas residuales significa simplemente que los errores típicos de medida son iguales entre las poblaciones.

La naturaleza compensatoria del DIF, a nivel de escala, se aborda en el contexto de la TRI en el procedimiento DFIT (Raju *et al*., 1995). Este aspecto no ha recibido demasiada atención en el resto de procedimientos basados en la TRI, ni en el contexto del AFC, a excepción del trabajo de McDonald (1999) sobre funcionamiento diferencial del test en el contexto del AFC. En la práctica, hay ítems que presentan DIF a favor de uno de los grupos mientras que otros ítems presentan DIF a favor del otro grupo, por lo que al sumar estas cantidades con signo opuesto pueden anular el funcionamiento diferencial del test. Tal y como apuntan Raju *et al.* (1995), esta información puede resultar muy útil a la hora de decidir qué hacer con los ítems que presentan DIF, especialmente cuando las razones por las que no hay equivalencia no están claras.

Tal y como señalan Flowers *et al.* (2002), no hay demasiados estudios que ofrezcan una comparación de sus resultados utilizando el AFC y algún procedimiento basado en la TRI para estudiar la equivalencia de un test. A continuación se describe brevemente el estado de la cuestión.

Raju, Laffitte y Byrne (2002) comparan un procedimiento basado en el AFC con el procedimiento DFIT en una escala de 10 ítems politómicos (5 categorías) con datos reales.

Encuentran un alto grado de acuerdo entre ambas técnicas, detectando dos ítems que presentan DIF, uno en ambas técnicas y otro únicamente en el AFC.

Tomás, González-Romá y Gómez (2000) comparan, con datos reales, el AFC (en entorno MACS) y la TRI como métodos alternativos para evaluar la equivalencia psicométrica en el contexto de la traducción de instrumentos de medida. Realizaron AFCs utilizando el modelo de medias latentes, y en el ámbito de la TRI utilizaron el MRG de Samejima, con una estrategia de comparación de modelos anidados basada en el test LR para comparar el valor de los parámetros en ambos grupos. La escala consta de 6 ítems politómicos (6 categorías) de los cuales, cuatro presentan funcionamiento diferencial con ambos métodos, uno con ninguno de ellos y uno presenta DIF utilizando la comparación de modelos mediante TRI pero no utilizando AFC. Concluyen que los resultados obtenidos por ambos métodos son muy similares.

El objetivo del estudio de Facteau y Craig (2001) consiste en determinar si un instrumento de evaluación sobre el rendimiento presenta equivalencia de medida entre cuatro grupos diferentes de evaluadores (uno mismo, compañeros, superiores y subordinados) en una escala de 8 dimensiones y 44 ítems con 5 categorías de respuesta. Para ello se utiliza el AFC multigrupo (forzando la igualdad de las cargas factoriales) y el procedimiento DFIT (estima los parámetros con el MRG de Samejima). Los resultados del AFC indican que el instrumento de evaluación fue invariante en los cuatro grupos de evaluadores, mientras que el procedimiento DFIT encuentra algún indicio de DIF, pero solo en tres ítems y de una magnitud trivial. En conjunto, los resultados apoyan la equivalencia de medida entre los grupos de evaluadores, lo que permite que sus puntuaciones en el test de rendimiento de los diversos grupos se comparen directamente.

El objetivo de Cooke, Kosson y Michie (2001) es comprobar la equivalencia métrica de un test que evalúa el grado de psicopatía entre caucásicos y afroamericanos, ya que todas las evidencias de validez del test encontradas hasta su investigación son estudios realizados con participantes caucásicos exclusivamente. Utilizaron el AFC para comprobar la unidimensionalidad del test y para probar la equivalencia de medida sin encontrar diferencias significativas entre el modelo sin restricciones y el modelo de igualdad de cargas factoriales, varianzas y errores. En el entorno TRI, para estimar los parámetros utilizan el MRG de Samejima (los ítems del test tienen tres opciones de respuesta), para calcular el DIF utilizan el test de razón de verosimilitud y el funcionamiento diferencial del test lo evalúan con el índice DTF del procedimiento DFIT. De los 20 ítems del test, 5 presentan DIF no uniforme, aunque estas diferencias entre ambos grupos en los ítems se anulan en la escala, al no haber indicios de funcionamiento diferencial del test.

Breithaupt y Zumbo (2002) estudian la equivalencia de medida con datos reales (6621 sujetos en una escala de 20 ítems dicotómicos) mediante comparación de modelos del AFC (forzando la igualdad de las cargas factoriales de cada ítem) y de la TRI (forzando la igualdad del parámetro de discriminación entre los grupos). Sus resultados apuntan a una falta de invarianza en las tres variables estudiadas (sexo, edad y grupo salud) cuando utilizan el AFC multigrupo pero no al basarse en las diferencias del parámetro de discriminación de la TRI, que consideran muy similar en todos los grupos estudiados. Atribuyen estas diferencias a un mejor funcionamiento del procedimiento basado en la TRI.

Maurer, Raju y Collins (1998) utilizan el AFC multigrupo y procedimiento DFIT para determinar el grado en que las evaluaciones realizadas por subordinados e iguales sobre la capacidad de trabajo en equipo de un directivo son directamente comparables. Para ello, utilizan una escala de ítems con cinco opciones de respuesta, que tuvieron que agrupar a 3 en el caso del análisis TRI por problemas de convergencia. En el AFC restringen la igualdad de las cargas factoriales de los 7 ítems entre ambos grupos sin haber diferencias significativas entre el modelo base y el modelo con restricciones. Los resultados del procedimiento DFIT también son acordes con la equivalencia de medida de la escala en ambos grupos.

También Reise *et al.* (1993) utilizan AFC y TRI para evaluar la equivalencia en este caso sobre una escala que mide el afecto negativo del estado de ánimo con 5 ítems politómicos (con 5 categorías de respuesta). En el AFC se fuerza la igualdad de cargas factoriales. En TRI, la estimación de parámetros se basa en el MRG con MULTILOG y analiza la equivalencia basándose en una medida de ajuste-persona llamada Zl (Drasgow, Levine y Williams, 1985). Bajo ambos procedimientos, sus resultados no son compatibles con un escenario de equivalencia completa entre los grupos, pero sí con la equivalencia parcial. Además, de los 5 ítems que constituyen la escala analizada dos presentan invarianza de medida utilizando ambos procedimientos y uno presenta DIF también bajo ambos métodos, mientras que los otros dos presentan DIF únicamente cuando se utiliza el procedimiento basado en la TRI. Los autores justifican estas diferencias en los resultados argumentando que el modelo de AFC ignora los parámetros *b* de la TRI, motivo por el cual los modelos basados en la TRI son más exigentes en los estudios de equivalencia.

El propósito del estudio de Scandura, Williams y Hamilton (2001) es realizar una investigación sustantiva de la medida en que una escala psicológica del comportamiento político en general en el ámbito de las organizaciones suscita respuestas equivalentes en muestras de Estados Unidos y de Oriente Medio. Analizan la equivalencia de medida de la escala (de 6 ítems con 5 alternativas de respuesta) en ambos grupos utilizando dos aproximaciones: el AFC multigrupo, en el que restringe la igualdad de las cargas factoriales de todos los ítems entre grupos; y el basado en la comparación de modelos basado en el test LR de la TRI, obligando en este caso a la igualdad de todos los parámetros de los ítems ($a$, $b_1$, $b_2$, $b_3$ y $b_4$) entre los grupos. En ambas aproximaciones sus resultados no son compatibles con un modelo de equivalencia total, por lo que utilizan modelos de equivalencia parcial con distintos resultados. En el caso de AFC basta con liberar las restricciones de igualdad de cargas factoriales de uno de los seis ítems de la escala para encontrar apoyo a la equivalencia parcial de medida, mientras que los resultados del modelo de comparación de modelos basado en la TRI no apoyan el establecimiento de equivalencia parcial de medida, esto es, los análisis TRI indican que la escala y sus ítems no son invariantes entre ambas culturas.

El propósito de la investigación de Kim, Kim y Kamphaus (2010) es establecer la equivalencia de medida entre sexos de un test de agresividad de 14 ítems politómicos (4 opciones de respuesta) para que puedan realizarse, con garantías de validez, las pertinentes comparaciones entre chicos y chicas. Utilizaron tanto procedimientos basados en el AFC (igualdad de cargas factoriales) como en la TRI (comparación de modelos basada en el test LR), rechazando desde ambas perspectivas la equivalencia total de medida entre sexos. En el ámbito de la equivalencia parcial encontraron que el AFC detectó más casos de ítems con DIF que el procedimiento de comparación de modelos de la TRI.

Dado que la literatura no ofrece datos concluyentes y puesto que es necesario un estudio de simulación para saber si realmente los ítems tienen o no DIF, se exponen a continuación los estudios con datos simulados que examinan el acuerdo o desacuerdo entre ambas aproximaciones basadas en AFC y en TRI.

Stark, Chernyshenko y Drasgow (2006) utilizan una estrategia común para identificar DIF en MACS y en TRI. Utilizando datos simulados de una escala de 15 ítems examinan simultáneamente las cargas factoriales y los interceptos en MACS y los parámetros de discriminación y localización utilizando el test de razón de verosimilitud en TRI, utilizando en ambos casos un modelo base y los valores *p* críticos de la corrección de Bonferroni. Comparan la eficacia de este procedimiento en varias condiciones: tipo y cantidad de DIF, tamaño muestral, número de categorías de respuesta y cantidad de impacto. Sus resultados indican que los procedimientos basados en MACS y TRI funcionaron bien y de manera similar en la mayoría de las condiciones experimentales. MACS funcionó peor en la condición de datos dicotómicos (como es de esperar) pero también en el caso de datos politómicos cuando los tamaños muestrales eran pequeños. Funcionó bien en las condiciones en las que se simuló DIF en los umbrales del ítem y su precisión no se vió afectada por el impacto.

El objetivo del estudio de Flowers, Raju y Oshima (2002) es comparar procedimientos MACS y TRI para evaluar la equivalencia de medida. Este estudio simula los datos de un test de 20 ítems de 5 alternativas de respuesta, utilizando el MRG de Samejima para examinar la ejecución de métodos basados en el AFC y en la TRI. Se utilizan dos procedimientos de AFC multigrupo para examinar la equivalencia de medida

entre grupo focal y grupo de referencia: forzando a la igualdad únicamente las cargas factoriales de los ítems y forzando además los interceptos. También se utiliza un índice basado en el procedimiento DFIT de la TRI, el índice NC-DIF (Raju *et al*., 1995) para examinar la equivalencia entre grupos. Los resultados indican que el procedimiento de igualdad de cargas factoriales del AFC identifica sucesivamente ítems que tienen diferencias en los parámetros *a*, pero no identifica ítems que tienen diferencias en los parámetros *b*. El procedimiento de igualdad de cargas factoriales e interceptos y el procedimiento NC-DIF identifican ítems que tiene diferencias en los parámetros *b*; sin embargo, no fueron sensibles a los ítems que tenían diferencias solo en los parámetros *a*. Cuando los grupos focal y de referencia tienen diferentes distribuciones en el rasgo (impacto), el procedimiento de invarianza escalar tiene una tasa de error Tipo II (detectar falsos negativos) baja, pero tiene una tasa de error Tipo I (detectar falsos positivos) alta. El procedimiento NC-DIF mostró tener tasas de errores tipo I y tipo II aceptables tanto en casos de impacto como de no impacto.

Meade y Lautenschlager (2004a) utilizan datos simulados para comparar la eficacia en el ámbito de MACS del AFC multigrupo (invarianza métrica e invarianza escalar) y de la comparación de modelos utilizando el test LR basado en la TRI. Los datos se simularon para reflejar las respuestas a una escala de seis ítems con cinco opciones de respuesta con tres tamaños muestrales: 150, 500 y 1000 y diversas condiciones de equivalencia y falta de equivalencia en los parámetros. Hipotetizan que los datos simulados para tener diferencias únicamente en el parámetro *b* serán detectados por el procedimiento basado en la TRI pero no por el AFC y que ambos procedimientos detectarán la falta de equivalencia en los datos con diferencias en el parámetro *a*. Sus resultados confirman únicamente la primera

hipótesis, ya que el AFC resultó inadecuado tanto para detectar diferencias en el parámetro *b* como para detectar diferencias en el parámetro *a*.

Meade y Lautenschlager (2004c) comparan, en el ámbito MACS, el AFC (invarianza métrica, escalar y de igualdad de las varianzas de los factores), el test de razón de verosimilitud y el procedimiento DFIT con datos simulados (6 ítems con 5 opciones de respuesta), encontrando, en todas las condiciones del estudio, que el test de razón de verosimilitud supera de forma constante en la evaluación de la equivalencia de medida tanto al AFC como al procedimiento DFIT basado en la TRI.

En líneas generales, las conclusiones de los estudios revisados no son directamente comparables ya que difieren en la forma de abordar el problema y en las condiciones de la investigación. Algunos autores utilizan modelos de invarianza métrica y otros de invarianza escalar en el AFC multigrupo y también son diferentes los procedimientos de equivalencia basados en la TRI. Los estudios basados en simulación también son escasos y tampoco ofrecen directrices concluyentes en esta cuestión.

# Sección II. ESTUDIO EMPÍRICO

# 1. OBJETIVOS

El principal objetivo de esta tesis es examinar los procedimientos considerados más relevantes para evaluar la equivalencia de medida de un test: el análisis factorial confirmatorio y dos procedimientos basados en la TRI, como son la comparación de modelos mediante el test de razón de verosimilitud y el procedimiento basado en el funcionamiento diferencial de ítems y tests (DFIT).

Para ello, se ha elegido trabajar con un constructo que tiene interés en diversos ámbitos de la Psicología (educativa, clínica y organizacional) y con un instrumento muy utilizado para medirlo, aplicado a una muestra representativa de la población escolar de la comunidad de Madrid: el Test de Impulsividad de Barrat. La literatura previa revela que en este test hay un factor dominante sobre el resto y que funciona mejor a nivel de escala global que de subescalas; por otro lado, los resultados son consistentes al mostrar que los chicos son más impulsivos que las chicas y que el constructo de impulsividad se configura de manera algo diferente en las distintas etapas evolutivas, pudiendo cambiar considerablemente de la preadolescencia a la adolescencia. Por tanto, este test resulta apropiado para examinar la equivalencia de medida con respecto a dos variables particularmente relevantes en relación con el constructo evaluado, como son el sexo y la edad y, por otro lado, su estructura factorial posibilita plantear el análisis tanto mediante modelos unidimensionales como multidimensionales.

Para lograr el objetivo central formulado, previamente es necesario evaluar la calidad métrica del test en la muestra del estudio. Para ello, se realiza un detallado análisis

de ítems y de la fiabilidad de la prueba desde la óptica de la teoría clásica de los tests y de la teoría de respuesta al ítem, examinando el ajuste del modelo de respuesta graduada de Samejima a los datos. Para obtener evidencias acerca de la validez de sus puntuaciones, se ha optado por examinar la estructura dimensional del test de Barrat mediante el análisis factorial confirmatorio.

## 2. MÉTODO

Los datos de esa investigación se han obtenido como parte de un proyecto más amplio, financiado por el Ministerio de Trabajo y Asunto Sociales (Proyecto RS/MS2001-16/01), cuyo objetivo último es proporcionar información que ayude a los responsables de los ámbitos educativos a comprender mejor las actitudes violentas y agresivas en preadolescentes y los adolescentes. Para obtener la información de interés se han aplicado, además del Test de Impulsividad de Barratt en cuyos datos se ha basado esta tesis doctoral, el Cuestionario de Agresión (AQ), la Escala de Agresión Directa e Indirecta (DIAS), la adaptación española del test STAXI y un cuestionario elaborado "ad hoc" para conocer los datos demográficos, hábitos y opiniones de los sujetos de estudio.

### 2.1. PARTICIPANTES

El estudio está dirigido a los escolares de la Comunidad de Madrid. De acuerdo con los objetivos de nuestra investigación hay dos rangos de edad de interés: los alumnos con edades comprendidas entre los 9 y 11 años y los de edades entre los 14 y 16 años. Los

cursos escolares correspondientes a dichas edades son: 4º y 5º de Primaria y 3º y 4º de Enseñanza Secundaria Obligatoria.

Según datos del Instituto Nacional de Estadística (INE), en el año 2003 hay en la Comunidad de Madrid 162.621 niños con edades comprendidas entre los 9 y 11 años, y 173.260 con edades entre los 14 y 16 años. A pesar de que estas cifras corresponden a datos de empadronamiento, cabe esperar, dada la edad, que la gran mayoría se encuentren escolarizados. Estas cifras hacen muy complicada la consideración del sujeto como unidad muestral, por lo que se opta por una unidad muestral mayor: el centro escolar.

Dada la naturaleza del estudio se desarrolla un muestreo probabilístico por conglomerados bietápico, de acuerdo con el siguiente procedimiento:

1) En la primera etapa, se construye el marco muestral. Para su elaboración se utiliza la información proporcionada en la Guía de Centros y Áreas Territoriales de la Consejería de Educación de la Comunidad de Madrid y en la base de datos del Instituto Nacional de Estadística "Sociedad y Educación". Los datos sobre las poblaciones y su número de habitantes se obtiene de la información del Instituto Nacional de Estadística sobre "Cifras de población de municipios por sexo".
2) A continuación se realiza un muestreo aleatorio estratificado para seleccionar los centros escolares, teniendo en cuenta dos variables de estratificación:
   - Tipo de centro escolar:
     - Público
     - Privado (incluidos centros concertados).
   - Tamaño de la población en la que está situado el centro escolar:
     - Hasta 10.000 habitantes,

- De 10.001 a 250.000 habitantes.
- Más de 250.000 habitantes.

3) En cada colegio seleccionado se elige de forma aleatoria las aulas a formar parte del estudio, aplicándose como máximo la batería de pruebas a un aula por cada nivel educativo.

En la Comunidad de Madrid, de los 1507 Centros Escolares que en el año 2003 imparten enseñanzas de Primaria y Secundaria, 1014 son públicos frente a 493 privados. En cuanto al *tamaño de la población* en la que está situado el centro escolar, en Madrid Capital, que es la única población con más de 250.000 habitantes hay 647 centros escolares de enseñanzas de Primaria y/o Secundaria, frente a los 717 centros de las poblaciones cuyo tamaño oscila entre 10.001 a 200.000 habitantes y los 142 centros de las poblaciones cuyo censo registra hasta 10.000 personas (ver Tabla 2.1).

Tabla 2.1. *Número de centros escolares de la Comunidad de Madrid en función del tipo de centro y del tamaño de la población*

| TIPO CENTRO | TAMAÑO DE LA POBLACIÓN | | | Total |
|---|---|---|---|---|
| | Hasta 10.000 hab. | De 10.000 a 250.000 hab. | Más de 250.000 hab. | |
| Público | 125 (8.3%) | 566 (37.6%) | 322 (21.4%) | 1014 (67.3%) |
| Privado | 17 (1.1%) | 151 (10.0%) | 325 (21.6%) | 493 (32.7%) |
| Total | 142 (9.4%) | 717 (47.6%) | 647 (43.0%) | 1507 |

La selección de centros dentro de cada estrato se realizó mediante muestreo aleatorio simple utilizando una tabla de números aleatorios.

El error de precisión, calculado con el programa informático SOTAM (Manzano, 1998), fue de ±6%, para un nivel de confianza del 95%.

La siguiente ficha técnica general resume los principales datos técnicos del muestreo realizado:

Tabla 2.2. *Ficha técnica del muestreo*

| **FICHA TÉCNICA** | |
|---|---|
| **Universo:** | Estudiantes de 4º y 5º de Primaria y 3º y 4º de ESO de la Comunidad de Madrid |
| **Diseño del muestreo:** | Muestreo por conglomerados polietápico |
| **Tamaño de la muestra:** | n = 2116 estudiantes |
| **Error de precisión:** | ± 6% |
| **Nivel de confianza:** | 95% |
| **Tipo de encuesta:** | Autoinforme, realizada en los centros escolares seleccionados |
| **Trabajo de campo:** | 2003-2004 |

El número final de centros seleccionados según el tipo de centro y el tamaño de la población en el que están situados se muestra en la siguiente tabla:

Tabla 2.3. *Número de centros escolares que participaron en la investigación en función del tipo de centro y del tamaño de la población*

| TIPO CENTRO | TAMAÑO DE LA POBLACIÓN | | | Total |
|---|---|---|---|---|
| | Hasta 10.000 hab. | De 10.000 a 250.000 hab. | Más de 250.000 hab. | |
| Público | 9 (33.3%) | 7 (25.9%) | 3 (11.1%) | 19 (70.3%) |
| Privado | 2 (7.4%) | 2 (7.4%) | 4 (14.8%) | 8 (29.6%) |
| Total | 11 (40.7%) | 9 (33.3%) | 7 (25.9%) | 27 |

Por otra parte, tal y como se puede observar en la Tabla 2.3, el porcentaje de centros escolares públicos de poblaciones pequeñas incluidos en la investigación es mucho mayor que el que le correspondería, teniendo en cuenta los datos relativos a la población de centros proporcionada en la Tabla 2.1. Esta diferencia está justificada por el fenómeno de los Centros Rurales Agrupados (C.R.A.), que son una respuesta organizativa para la gestión educativa (de recursos materiales y humanos) de pequeñas escuelas rurales pertenecientes a un mismo entorno físico, social y natural. Así, en la Comunidad de Madrid hay ocho Centros Rurales Agrupados, que dan lugar a 47 escuelas de diferentes municipios. El C.R.A. seleccionado aleatoriamente en este estudio fue "Amigos de la Paz" cuya dirección y gestión se organiza desde el colegio de Anchuelo y comprende las escuelas de Anchuelo, Corpa, Pezuela de las Torres, Santorcaz, Los Santos de Humosa y Valverde de Alcalá.

Al considerar como centro escolar el C.R.A. y no la escuela rural es necesario aplicar una corrección en el número de centros escolares públicos situados en poblaciones de hasta 10.000 habitantes, tanto en el ámbito muestral como en el poblacional (ver Tablas 2.4 y 2.5).

Tabla 2.4. *Número corregido de centros escolares de la Comunidad de Madrid en función del tipo de centro y del tamaño de la población*

| TIPO CENTRO | TAMAÑO DE LA POBLACIÓN | | | Total |
|---|---|---|---|---|
| | Hasta 10.000 hab. | De 10.000 a 250.000 hab. | Más de 250.000 hab. | |
| Público | 86 (5.9%) | 566 (38.6%) | 322 (21.9%) | 974 (66.4%) |
| Privado | 17 (1.2%) | 151 (10.3%) | 325 (22.1%) | 493 (33.6%) |
| Total | 103 (7.0%) | 717 (48.9%) | 647 (44.1%) | 1467 |

Tabla 2.5. *Número corregido de centros escolares que participaron en la investigación en función del tipo de centro y del tamaño de la población*

| TIPO CENTRO | TAMAÑO DE LA POBLACIÓN | | | Total |
|---|---|---|---|---|
| | Hasta 10.000 hab. | De 10.000 a 250.000 hab. | Más de 250.000 hab. | |
| Público | 4 (18.2%) | 7 (31.8%) | 3 (13.6%) | 14 (63.6%) |
| Privado | 2 (9.1%) | 2 (9.1%) | 4 (18.2%) | 8 (36.4%) |
| Total | 6 (27.3%) | 9 (40.9%) | 7 (31.8%) | 22 |

De los centros escolares en los que se ha llevado a cabo la aplicación de los tests siete están situados en Madrid capital y el resto en pueblos de la periferia. En las Figuras 2.1. y 2.2 se muestra el mapa de Madrid en el que se han recuadrado los centros escolares seleccionados en la investigación situados en Madrid capital (ver Figura 2.1) y en el resto de la comunidad (ver Figura 2.2).

Centros situados en Madrid Capital:

- Emilia Pardo Bazán (distrito Centro)

- Santa Isabel (distrito Centro)
- Ntra. Sra. de Loreto (distrito Salamanca)
- Jaime Vera (distrito Tetuán)
- Divino Corazón (distrito Tetuán)
- Ntra. Sra de los Dolores (distrito Carabanchel)
- Dr. Conde Arruga (distrito Moratalaz)


Figura 2.1. Mapa de Madrid capital con los distritos seleccionados.

Centros situados en la periferia:

- San José de Calasanz (Getafe)
- Antonio López García (Getafe)
- Divina Pastora (Getafe)
- La Laguna (Parla)

- Camilo José Cela (Pozuelo de Alarcón)
- Isabel la Católica (Pinto)
- Príncipe D. Felipe (Boadilla del Monte)
- Ventura Rodríguez (Ciempozuelos)
- Antonio Machado (Meco)
- Vicente Aleixandre (Miraflores de la Sierra)
- San Pablo (Miraflores de la Sierra)
- Santa Elena (Villarejo de Salvanés)
- SIES de Griñón (Griñón)
- Santísima Trinidad (Collado Villalba)
- CRA de Anchuelo (Centro Rural Agrupado que comprende los municipios de Anchuelo, Corpa, Pezuela de las Torres, Santorcaz, Los Santos de la Humosa y Valverde de Alcalá)

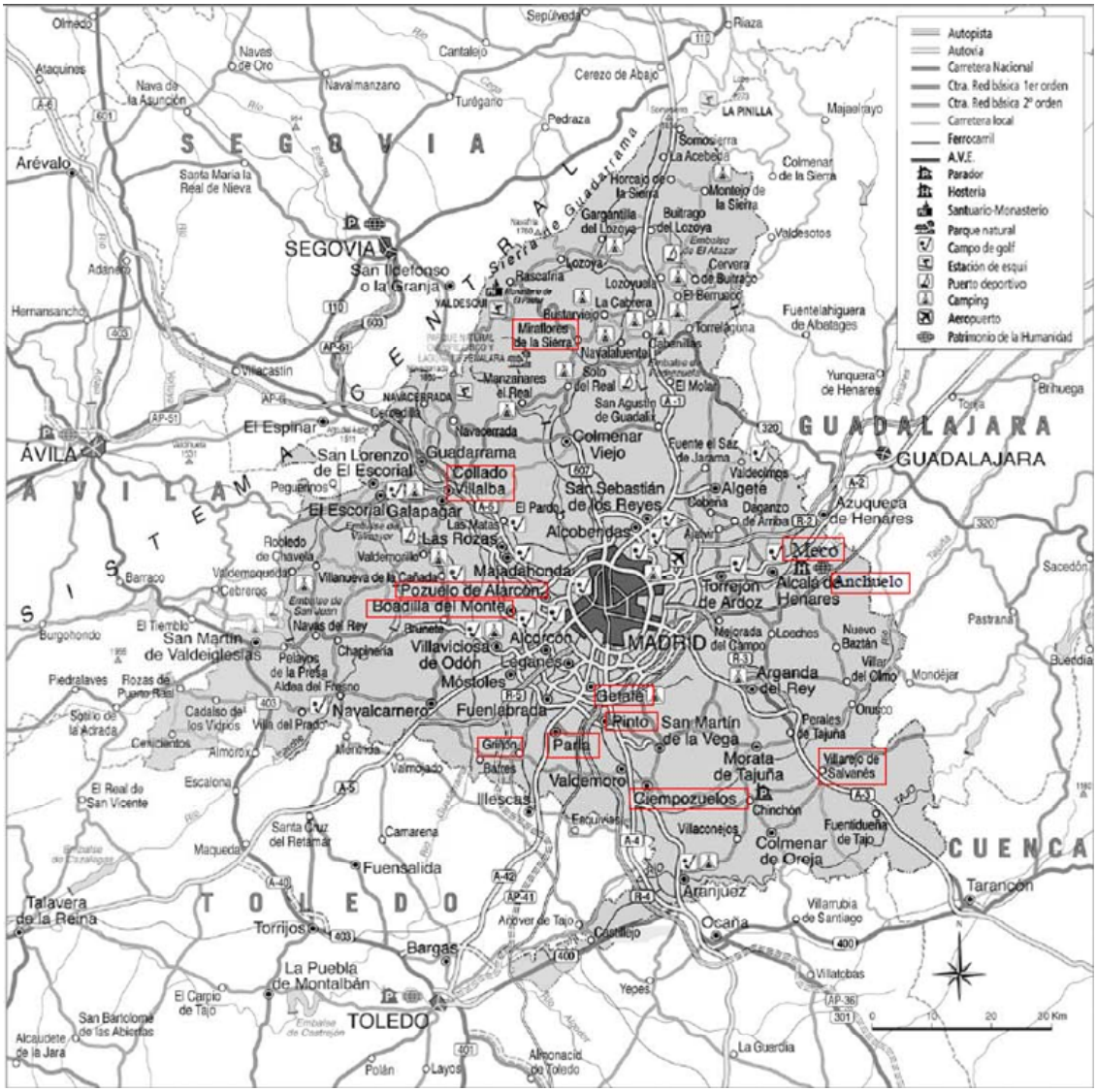



Figura 2.2. Mapa de Madrid con las localidades seleccionadas situadas en la periferia.

El procedimiento de asignación de tamaños muestrales a los diferentes estratos ha sido no proporcional. Esta decisión está motivada por las diferencias en el número de elementos de cada estrato en las poblaciones, ya sea teniendo en cuenta la variable de estratificación *tipo de centro* o *tamaño de la población en la que está situado el centro escolar*. En caso de utilizar asignación proporcional el tamaño muestral de los estratos se determina con la fórmula: $n_i = \frac{n \times N_i}{N}$, lo que nos hubiera llevado, por ejemplo, a una muestra de 0.25 centros en el caso de centros públicos situados en poblaciones de menos de 10000 habitantes.

Dado que se ha realizado una asignación no proporcional de la muestra a los diferentes estratos es conveniente realizar una ponderación de los datos (ver por ejemplo Kalton, 1983). Para ello, se determina el peso asignado a cada centro según la siguiente fórmula:

$$w_{ij} = \frac{N_i}{N} \cdot \frac{n}{n_i}$$

donde:

$w_{ij}$ es el peso del centro *j* del estrato *i* (idéntico para todos los centros del mismo estrato)

$N_i$ es el número de unidades en la población en el estrato i-ésimo

$N$ es el tamaño de la población

$n_i$ es el tamaño de la muestra en el estrato i-ésimo

$n$ es el tamaño de la muestra.

Los pesos asignados a los diferentes estratos tras la aplicación de la corrección por ponderación se muestran en la Tabla 2.6. Dicha ponderación se utilizará en los análisis relativos a las diferencias en impulsividad entre hombres y mujeres y en los dos rangos de edades estudiadas.

Tabla 2.6. *Pesos asignados a los diferentes centros escolares en función del tipo de centro y del tamaño de la población donde está situado*

| TIPO CENTRO | TAMAÑO DE LA POBLACIÓN | | |
|---|---|---|---|
| | Hasta 10.000 hab. | De 10.000 a 250.000 hab. | Más de 250.000 hab. |
| Público | 0.32 | 1.21 | 1.61 |
| Privado | 0.13 | 1.13 | 1.22 |

La población objetivo del estudio son estudiantes tanto de Primaria (4º y 5º) como de Secundaria (3º y 4º) de la Comunidad de Madrid. Dado que los tests y cuestionarios se han administrado a todos los alumnos de los grupos seleccionados, hay sujetos que se han eliminado de la muestra final, al no estar su edad en el rango deseado. De este modo, se han excluido a1 final 157 participantes por este motivo: 1 sujeto con 8 años, 18 sujetos con 12 años, 2 sujetos con 13 años, 106 sujetos con 17 años y 30 sujetos con 18 años.

La muestra definitiva está constituida por 2116 estudiantes. De ellos, aproximadamente la mitad pertenecen a un colegio público (49.5%) y la otra mitad a uno privado (50.5%). En cuanto al tamaño de la población, el 17.8% acuden a centros de hasta 10000 habitantes, frente al 43.2% de las poblaciones de tamaño intermedio y el 39% de Madrid Capital (ver Tabla 2.7).

Tabla 2.7. *Distribución de los sujetos que forman parte de la muestra seleccionada, en función del tipo de centro y del tamaño de la población en la que está situado*

| TIPO CENTRO | TAMAÑO DE LA POBLACIÓN | | | Total |
|---|---|---|---|---|
| | Hasta 10.000 hab. | De 10.000 a 250.000 hab. | Más de 250.000 hab. | |
| Público | 290 | 600 | 158 | 1048 |
| Privado | 87 | 314 | 667 | 1068 |
| Total | 377 | 914 | 825 | 2116 |

En cuanto a las características demográficas, el tamaño de la muestra por grupo de edad y sexo está bastante compensado, tal y como se puede apreciar en la Tabla 2.8.

Tabla 2.8. *Distribución de la muestra por edad y sexo*

| SEXO | Grupos de edad | | Total |
|---|---|---|---|
| | 9 a 12 años | 14 a 16 años | |
| Hombre | 507 (24.4%) | 409 (19.6%) | 916 (44.0%) |
| Mujer | 592 (28.4%) | 574 (27.6%) | 1166 (56.0%) |
| Total | 1099 (52.8%) | 983 (47.2%) | 2082 (100%) |

## 2.2. INSTRUMENTO

La versión original del Test de Impulsividad de Barratt (Barratt Impulsiveness Scale, BIS) fue elaborada por Barratt en 1959 bajo la denominación de BIS-1. Esta primera versión, de 80 ítems, constituye la primera escala de impulsividad que no forma parte de un inventario de personalidad. A partir de esta escala se han realizado continuas modificaciones, lo que ha dado lugar a versiones de menor longitud hasta llegar a la actual,

que contiene 30 ítems. Esta última versión ha sido desarrollada por Patton, Stanford y Barratt (1995) bajo la denominación de BIS-11. Cada ítem del test es una escala de categorías de frecuencia, en la que el participante responde al enunciado eligiendo una de sus cuatro alternativas de respuesta: "nunca o casi nunca", "algunas veces", "bastantes veces" y "siempre o casi siempre".

El test BIS-11 traducido al castellano por Oquendo, Baca-García, Graver, Morales, Montalvan y Mann (2001) es un cuestionario para adultos, por lo que reformulamos algunos de los ítems que no eran comprensibles para la población objeto de estudio. En concreto, se cambió el enunciado de 19 de los 30 ítems que componen la prueba. Por ejemplo, el ítem 5 de la versión original se formuló como: "I plan trips well ahead of time", y fue traducido por Oquendo *et al.* (2001) por "Planifico mis viajes con antelación". Dado que en el rango de edad estudiado no tiene sentido la pregunta se tuvo que reformular para que fuera aplicable a la muestra objeto de estudio, quedando de la siguiente manera: "Hago mis planes con mucho tiempo".

Se comprobó la correcta comprensión de los ítems reformulados mediante un estudio piloto en el que se preguntó a una muestra incidental de 20 estudiantes de 9 a 11 años sobre el significado de cada ítem, encontrando que entendían de manera adecuada los enunciados. En el Anexo 1 se incluyen los 30 ítems de esta adaptación del BIS-11.

Según Barratt, se pueden precisar 3 tipos de impulsividad: la impulsividad motora, la impulsividad cognitiva y la improvisación/ausencia de planificación (Patton, Standford y Barratt, 1995). La puntuación total se obtiene sumando las puntuaciones obtenidas.

La impulsividad motora (IM) se define como actuar sin pensar, dejándose llevar por el ímpetu del momento. Está definida por los ítems 2, 6, 9, 12, 15, 18, 23, 26 y 29. La impulsividad no-planificadora (INP) o improvisación/ausencia de planificación se caracteriza por la tendencia a no planificar mostrando un mayor interés en el presente que en el futuro. Está caracterizada por los ítems 1, 3, 5, 8, 11, 14, 17, 22, 25, 28, 30. La impulsividad cognitivo-atencional (ICA) implica una propensión a tomar decisiones cognitivas rápidas. Tiene que ver con la rapidez de los pensamientos y con la atención en el sentido de no ser capaz de focalizar la atención en la tarea que se está ejecutando. Está definida por los ítems 4, 7, 10, 13, 16, 19, 20, 21, 24 y 27.

Esta descripción sobre la dimensionalidad del BIS no está exenta de controversia. A pesar del escaso número de investigaciones sobre las propiedades de la escala, desde 1985, fecha en la que Barratt identificó 3 sustratos principales de impulsividad existe desacuerdo en cuanto a la estructura factorial tanto de la escala como de sus diversas adaptaciones habiendo trabajos que arrojan resultados contradictorios (Bayle *et al.*, 2000; Chahin, Cosi, Lorenzo-Seva y Vigil-Colet, 2010; Fossati, Barratt, Acquarini y Di Ceglie, 2002; Fossatti, Di Ceglie, Acquarini y Barratt, 2001; Someya *et al.*, 2001). La mayoría de estos trabajos contempla una estructura de seis factores de primer orden y tres factores de segundo orden, aunque son pocos los que ponen a prueba esta estructura. Nuestra adaptación de la escala fue objeto de estudio en un trabajo anterior (Recio *et al*., 2004).

El BIS es un test diseñado para medir impulsividad, por lo que se utiliza en el ámbito clínico como parte del método diagnóstico para detectar enfermedades relacionadas con altos niveles de impulsividad (trastornos bipolares, alcoholismo y abuso de sustancias, hiperactividad infantil, trastornos obsesivo-compulsivos...), en el ámbito educativo para

discriminar el origen de conductas inapropiadas en el aula como falta de atención o desobediencia, e incluso en el ámbito laboral, ya que hay selecciones de personal que incluyen baterías completas de tests de personalidad, en los que se puede evaluar este rasgo.

En el ámbito clínico se intentarán detectar casos "anormales" de impulsividad, ya sea por exceso o por defecto. Esto es, la mayoría de los sujetos presentarán puntuaciones no patológicas de impulsividad y el test se utilizará en muchas ocasiones para detectar personas con este problema. En selección de personal, cabe la posibilidad, además, de que el perfil del puesto requiera personas con niveles relativamente altos o bajos del rasgo, pero sin llegar a considerarse patológicos (es preferible por ejemplo, una persona con bajos niveles de impulsividad para tareas que requieren atención sostenida en el tiempo).

En el ámbito educativo se puede utilizar para explicar ciertas conductas del alumno en el aula. En el caso concreto de niños y adolescentes, la impulsividad está implicada en problemas de lectura, el trastorno por hiperactividad y déficit de atención, etc., que, a su vez, generan problemas de aprendizaje y fracaso escolar (Harmon-Jones, Barratt y Wigg, 1997). Según Barratt (1994), los sujetos impulsivos tienen más problemas para aprender que los sujetos con bajos niveles de impulsividad, lo que implica que la impulsividad podría estar relacionada con el fracaso escolar.

Ya en la vida adulta la impulsividad se considera un aspecto clave en la evaluación del riesgo de ejercer conductas violentas y suele estar relacionada con conductas de juego patológico, consumo de sustancias ilícitas y otros comportamientos incontrolados (Barratt, 1994; Hart y Dempster, 1997). Algunos autores van más allá en sus afirmaciones,

considerando a la impulsividad como el mejor predictor de conducta antisocial/delictiva en la edad adulta (Knorring y Ekselius, 1998; Tremblay, Pihl, Vitaro y Dobkin, 1994).

No es de extrañar, por tanto, la atención que recibe en la literatura científica el rasgo de impulsividad (se puede consultar una revisión exhaustiva en Arce y Santisteban, 2006). En general, las investigaciones han sido consistentes al encontrar que los hombres tienen unos mayores niveles de impulsividad que las mujeres (Chapple y Johnson, 2007) y que el constructo de impulsividad se configura de manera ligeramente diferente en las distintas etapas evolutivas, pudiendo cambiar considerablemente de la preadolescencia a la adolescencia.

## 2.3. RECOGIDA DE DATOS

Una vez obtenido el permiso del centro y concertada una cita, se administró el test a los alumnos durante el periodo de clase. Los datos se recogieron en el curso académico 2003/2004. La prueba fue administrada en los centros escolares por encuestadores entrenados, dentro del aula y en el horario académico de los estudiantes, siempre con el margen de tiempo necesario para permitirles contestar con total libertad y sin premuras de tiempo. En las instrucciones de la prueba (ver anexo 2), que fueron idénticas en todas las aplicaciones, se hizo hincapié en la importancia que tiene responder con total sinceridad, así como en el anonimato de las respuestas.

## 2.4. ANÁLISIS ESTADÍSTICOS

### 2.4.1. PROPIEDADES PSICOMÉTRICAS DEL TEST BIS

Para comprobar que el test BIS reúne los requisitos de calidad métrica necesarios para realizar las diferentes pruebas de equivalencia de medida se realizan análisis factoriales, análisis clásicos y análisis basados en la TRI.

En primer lugar se lleva a cabo un estudio de la dimensionalidad del test mediante comparación de modelos, teniendo en cuenta una estrategia de validación cruzada. A continuación se realiza un análisis de las propiedades psicométricas de los ítems basado en la TCT y en la TRI, y ambos procedimientos, además de métodos factoriales, se utilizan para abordar el estudio de la fiabilidad de la escala completa y sus subescalas. Por último, se evalúa el ajuste a los datos del modelo TRI elegido, con información de tipo gráfico y estadístico, también desde una estrategia de validación cruzada y una vez comprobado el requisito de unidimensionalidad necesaria.

En relación a los valores perdidos se puede optar por eliminar estos casos o por asignarles un valor determinado. Existen varios métodos para sustituir por una puntuación los valores perdidos, como su sustitución por la media del grupo o el método de imputación por máxima verosimilitud. De todas formas, la imputación de valores debe realizarse de la manera más cuidadosa y controlada posible porque los valores perdidos serán reemplazados por otros valores que serán tratados como datos reales observados. Según Jöreskog y Sörbom (1996) es preferible evitar utilizar variables con datos imputados en ecuaciones estructurales con LISREL. Si se incluyen es probable que la imputación

afecte al resultado de los análisis. Esto debería comprobarse comparando los resultados con y sin imputación de valores perdidos.

En la presente investigación se eliminan los registros perdidos en todos los análisis factoriales confirmatorios para no introducir artificios en la investigación, que en este caso, además, podrían ser de diferente magnitud en los análisis de equivalencia para cada grupo analizado. El tamaño muestral después de su eliminación resulta apropiado, ya que el ratio de número de sujetos en relación con el número de ítems fue mayor que 20:1 (Bollen, 1989). La muestra final analizada consta de 1690 participantes.

#### *2.4.1.1. Validez*

Se estudia la dimensionalidad del BIS como evidencia de validez, dado que todavía hay bastante controversia respecto a la estructura factorial del Test de Impulsividad de Barratt (ver apartado 2.3), realizando una comparativa del ajuste de tres estructuras factoriales (unifactorial, bifactorial y trifactorial).

Los modelos se ponen a prueba mediante AFC con el programa LISREL 8.54. Dado que las variables observadas de los modelos son ordinales (las respuestas a los ítems del test), se analiza la matriz de correlaciones policóricas (calculada con PRELIS 2.30) utilizando como método de extracción el método de mínimos cuadrados ponderados robusto (DWLS). Este procedimiento proporciona estimaciones correctas de los errores en muestras grandes (Joreskog, 1994, 2002).

En primer lugar se comprueba que los índices globales de ajuste de las tres estructuras son apropiados. Para la interpretación del ajuste del modelo se utiliza como índice de ajuste absoluto GFI, considerándose indicadores de buen ajuste los valores superiores a .90 (Bollen y Long, 1993; Byrne, 2001), y RMSEA, siendo los valores de hasta .08 indicativos de un ajuste razonable y los valores mayores de .10 una explicación inadecuada de los datos (Browne y Cudeck, 1993). Como índices de ajuste incrementales se utilizan el NNFI y el CFI, considerándose apropiados los valores superiores a .90 (Bentler, 1990). Además, se utiliza el ECVI, que fue propuesto por Browne y Cudeck (1989) para comparar modelos alternativos cuando sólo se utiliza una muestra. A medida que ECVI es más pequeño, o no varía entre los distintos modelos examinados, se entiende que el modelo se mantiene estable en la población.

El estadístico $\chi^2$ de bondad de ajuste se utiliza para comparar el ajuste de las diversas estructuras factoriales puestas a prueba, calculando las diferencias en los valores del estadístico $\chi^2$ entre los modelos para determinar si hay diferencias significativas entre ellos (se determina significativa la diferencia utilizando la diferencia en grados de libertad ($\Delta$g.l.) a un nivel $\alpha$ especificado a priori). También se utiliza para este mismo propósito la diferencia entre los valores del índice CFI, siendo relevantes las diferencias superiores a .01 siguiendo los criterios de Cheung y Rensvold (2002).

Para valorar la capacidad de generalización del modelo se lleva a cabo un procedimiento de validación cruzada, dividiendo la muestra aleatoriamente en dos partes para validar los resultados. Así, la Muestra 1 sirve como muestra de calibración, evaluándose el modelo inicialmente propuesto, y la validez de su estructura se comprueba con la muestra 2 o muestra de validación.

La validación cruzada se puede dar en varios grados (Bentler, 1980; MacCallum, Rosnowski, Mar y Reith, 1994). Bentler propuso una aproximación, denominada validación cruzada débil, que se limita a re-estimar todos los parámetros del modelo en una muestra independiente y que no está exenta de críticas. MacCallum *et al.* (1994) consideran que esta aproximación puede resultar útil como un mecanismo para evaluar la replicabilidad y estabilidad de las soluciones en términos de estimación de parámetros y bondad de ajuste, pero que no es una verdadera validación cruzada, porque el análisis de la muestra de validación no depende en ninguna medida de los resultados del análisis de la muestra de calibración.

Bentler (1980) también propone estrategias de validación cruzada moderada y fuerte, que implican restricciones sobre la igualdad de conjuntos de parámetros (igualdad de cargas factoriales, de covarianzas factoriales y de unicidad). MacCallum *et al.* (1994) identifican un procedimiento de validación cruzada jerárquico dependiendo de los parámetros que se fuerzan a ser iguales entre ambas muestras, considerando adecuada la validación cruzada cuando se encuentra igualdad entre muestras de calibración y validación en la estructura factorial, las cargas factoriales de los ítems y las covarianzas entre los factores. Éstas son las premisas que vamos a considerar: invarianza de configuración (modelo 1), invarianza métrica (modelo 2), invarianza de las covarianzas entre los factores (modelo 6), e invarianza de las varianzas error (modelo 4). La lógica del análisis es similar a la demostrada por Jöreskog y Sörbom (1996) para evaluar la equivalencia entre grupos.

En el caso de que ambas muestras resulten equivalentes, se unificarán para realizar los análisis de equivalencia en sexo y edad.

#### *2.4.1.2. Análisis de ítems*

Se evalúa la calidad de los ítems que conforman la prueba desde el modelo clásico y desde la TRI. Los análisis clásicos incluyen estadísticos descriptivos de cada ítem, así como la discriminación de los ítems mediante correlación ítem-test. Los análisis basados en la TRI incluyen la estimación de los parámetros de cada ítem con el MRG de Samejima.

#### *2.4.1.3. Fiabilidad*

En el estudio de la fiabilidad se ha complementado la utilización del coeficiente α (procedimiento clásico) con la función de información del test (procedimiento basado en TRI) y con otros índices factoriales, basados en el análisis de la matriz de correlaciones policórica –matriz de análisis apropiada en el caso de variables ordinales–, el coeficiente α ordinal, el coeficiente θ y el cálculo de la fiabilidad basado en el AFC.

El motivo de realizar otras estimaciones de fiabilidad es que, a pesar de que el coeficiente alpha de Cronbach (1951) es el indicador sobre la calidad métrica de un test del que más frecuentemente se ha informado en la literatura en ciencias sociales (Zumbo y Rupp, 2004), cada vez son más los autores que consideran que en la actualidad existen procedimientos que han demostrado ser más eficaces para este cometido (Sijtsma, 2009b) por lo que recomiendan restringir su uso o complementarlo con algún otro indicador de la

fiabilidad, ya sea basado en el análisis factorial, en la TRI o en la teoría de la generalizabilidad. Algunas críticas aluden a errores sistemáticos que sobreestiman o subestiman este coeficiente en función de la dimensionalidad del test (Cortina, 1993; Schmitt, 1996), de manera que su interpretación como límite inferior de la fiabilidad o en términos de consistencia interna, sobre todo en escalas ordinales, está siendo cuestionada (Bentler, 2009; Green y Yang, 2009; Sijtsma, 2009a).

El coeficiente alfa se estima sobre la matriz de correlaciones producto-momento de Pearson o de covarianzas que asume la naturaleza continua de las variables, por lo que su aplicación podría no ser correcta cuando la naturaleza de la escala de respuesta es ordinal, en especial en ítems con pocas opciones de respuesta (Elosua y Zumbo, 2008). En este sentido, varios estudios han mostrado que la utilización del coeficiente alfa sobre escalas de respuesta Likert con menos de 5 categorías de respuesta produce un decremento espurio en su magnitud (Lozano, García Cueto y Muñiz, 2008; Weng, 2004; Zumbo, Gadermann y Zeisser, 2007).

#### *2.4.1.4. Ajuste del modelo*

Se utiliza un modelo de la TRI unidimensional a pesar de que el test tiene una estructura trifactorial, por los siguientes motivos (Bolt, Hare, Vitale y Newman, 2004): (1) La unidimensionalidad en sentido estricto no es necesaria para beneficiarse de los beneficios de la utilización de la teoría de respuesta al ítem (Harrison, 1986; Smith y Reise, 1998), siempre y cuando haya un factor dominante sobre el resto. (2) La multidimensionalidad tiene el potencial de producir DIF cuando se aplica un modelo unidimensional. Esto es, un ítem que funciona igual en todos los grupos puede parecer que

funciona de manera diferencial si los grupos tienen distribuciones diferentes en el factor secundario. A este respecto, la dimensión secundaria no impide necesariamente el análisis DIF pero puede contribuir a proporcionar una interpretación del funcionamiento diferencial cuando éste se da. (3) Si la utilización práctica del test supone una puntuación total en el mismo, se considera más informativo estudiar la ejecución en el rasgo global subyacente que en los múltiples rasgos.

El BIS tiene un factor dominante sobre el resto y funciona mejor a nivel de test global que a nivel de subtests, por lo que es susceptible de análisis tanto mediante modelos unidimensionales como multidimensionales.

Antes de utilizar un modelo de la TRI unidimensional es necesario evaluar el ajuste, esto es, si las pruebas poseen la unidimensionalidad necesaria. Se atiende a las evidencias encontradas en los análisis factoriales confirmatorios realizados, además de utilizar el análisis factorial de componentes principales para evaluar la unidimensionalidad de la escala completa BIS y de sus subescalas, evaluándose dos aspectos: Primero, siguiendo las recomendaciones de Reckase (1979), el porcentaje de varianza explicado por el primer factor debe ser mayor que el 20%. Aunque este es el porcentaje mínimo de varianza explicada para la identificación de unidimensionalidad, Drasgow y Parsons (1983) demuestran que violaciones sustanciales de la unidimensionalidad no justifican necesariamente la utilización de un modelo multidimensional de TRI, ya que los modelos unidimensionales son robustos a la violación de este supuesto. Segundo, se examina el gráfico de sedimentación de los autovalores para determinar si hay un primer factor dominante.

Los parámetros de ítems y personas para cada grupo se estiman mediante el MRG de Samejima (1969), utilizando el programa MULTILOG. Para ítems puntuados en cuatro categorías, este modelo caracteriza cada ítem de acuerdo a cuatro parámetros: $b_1$, $b_2$ y $b_3$ son parámetros relativos a los umbrales del ítem o su localización, esto es, aluden al nivel de $\theta$ necesario para adscribirse en una categoría superior, mientras que $a$ es el parámetro de discriminación.

Para evaluar la adecuación del MRG a los datos considerados en este estudio, se recurre nuevamente a la validación cruzada, tal y como recomiendan Drasgow *et al.* (1995) realizando la estimación de los parámetros de los ítems con la muestra de calibración y estimando el nivel de aptitud de cada sujeto con la muestra de validación.

El ajuste al modelo se evalúa con el programa MODFIT (Stark, 2001) que proporciona información de tipo gráfica y estadística del ajuste al modelo especificado. En la valoración del ajuste gráfico se tiene en cuenta la correspondencia entre la curva teórica y la curva empírica de cada alternativa del ítem.

Respecto al uso del estadístico de razón de verosimilitud, puede resultar inadecuado utilizar el test de significación para evaluar el ajuste del modelo porque son muy sensibles al tamaño muestral y en una muestra bastante grande cualquier modelo de TRI sería rechazado (Drasgow, Levine, Tsien, Williams y Mead, 1995). Además de este problema, el estadístico de razón de verosimilitud presenta una potencia estadística baja con tamaños muestrales pequeños y una tasa de error tipo I muy elevada cuando el tamaño muestral es grande (Hambleton y Swaminathan, 1985; Hambleton, Swaminathan y Roger, 1991; López-Pina e Hidalgo, 1996; Orlando y Thissen, 2000). Por este motivo, el estadístico de

razón de verosimilitud se utiliza únicamente en este trabajo en la comparación de modelos y no para la valoración del ajuste.

En la valoración del ajuste estadístico se utilizan tres tipos de índices $\chi^2$ que evalúan el ajuste al MRG con respecto a las frecuencias conjuntas de las puntuaciones del ítem de primer-orden, segundo-orden y tercer-orden, respectivamente (para ver detalles, consultar Drasgow *et al.*, 1995). Esencialmente, los índices de primer orden evalúan si las probabilidades de las puntuaciones del modelo implícito según los niveles del rasgo, son consistentes con las probabilidades empíricas observadas para los ítems individuales. Los índices de segundo orden se computan comparando la probabilidad esperada y observada para las opciones específicas de los dos ítems (con una tabla de contingencia se compara la probabilidad observada y esperada de elegir la opción 1 en el ítem 1 y la opción 2 en el ítem 2, etc.). Los índices de tercer orden se calculan de manera similar siendo la tabla de contingencia en este caso de tres vías. Los índices de segundo-orden y tercer-orden son sensibles también a la dependencia local entre las puntuaciones de los ítems de dos en dos y de tres en tres. Siguiendo las recomendaciones de Drasgow *et al.* (1995) se considera que el ajuste al modelo es bueno si el índice de $\chi^2$ dividido entre los grados de libertad es igual o menor que tres.

### 2.4.2. IMPACTO

Para evaluar el impacto se evalúan las diferencias reales en impulsividad según la escala BIS y sus subescalas en las dos variables estudiadas (sexo y edad) realizando un análisis multivariante de la varianza. En estos análisis se han asignado pesos diferentes a los sujetos de la muestra, en función de las variables de estratificación, (ver Tabla 2.6.) con

el fin de otorgar una mayor (o menor) importancia relativa a algunas unidades muestrales en el análisis estadístico. Esta ponderación está motivada por el hecho de haber utilizado un procedimiento de asignación no proporcional de la muestra a los diferentes estratos (ver explicación en el apartado 2.1).

### 2.4.3. INVARIANZA MEDIANTE AFC MULTIGRUPO

En un AFC multigrupo con datos ordinales, la matriz de covarianzas tiene un significado distinto del análisis con un único grupo, ya que se trata de una matriz de correlación policórica escalada (Jöreskog y Sörbom, 1996). Primero, con PRELIS 2 se estiman las correlaciones policóricas y las medias y desviaciones típicas bajo los mismos umbrales, y después se "escala" la matriz de correlación a una matriz de covarianzas utilizando las desviaciones típicas estimadas. Esta matriz resultante es la que se utiliza en análisis multigrupo con variables ordinales (ver Jöreskog y Aish, 1996 para un ejemplo y Jöreskog y Sörbom, 1996 para consultar el desarrollo matemático).

Se utiliza un procedimiento en varios pasos. En primer lugar, se establece un modelo base, en el que la hipótesis a contrastar es que el patrón de cargas factoriales sea el mismo en los dos grupos, lo que se denomina en la literatura invarianza de configuración (Horn, McArdle y Mason, 1983). Se ejecuta el programa LISREL (8.54) que proporciona un valor de $\chi^2$, además de otros índices, para evaluar el ajuste del modelo en ambos grupos. Si el ajuste es adecuado, el modelo especificado sirve como modelo base de comparación.

Después se comprueba la invarianza métrica entre grupos. Para ello, se ejecuta nuevamente el programa con la restricción de igualdad de parámetros λ (cargas factoriales) en los dos grupos. Para comparar el modelo resultante con el modelo obtenido en el paso 1 (modelo base) se halla la diferencia entre los $\chi^2$ de los pasos 1 y 2 ($\Delta\chi^2$). La significación estadística de esta diferencia se valora utilizando la diferencia en grados de libertad (Δg.l.) a un nivel α especificado a priori. También se tiene en cuenta la diferencia entre ambos modelos anidados en el índice comparativo de Bentler (CFI), que no debe ser superior a .01, según los criterios de Cheung y Rensvold (2002) y Chen (2007).

Si no se encuentran diferencias significativas, esto significa que un modelo en el que los factores de carga se fuerzan a ser iguales en los dos grupos se ajusta a los datos tan bien como un modelo en el que los factores de carga se estiman de forma libre. Por tanto, los factores de carga son invariantes en los grupos, lo que apoya la hipótesis de equivalencia métrica. Si se encuentran diferencias significativas, hay que localizar los ítems que provocan esta falta de equivalencia, en el marco de la equivalencia parcial de medida (Byrne, 1989), dejando libres de restricción, uno a uno, a los ítems cuyo desajuste sea mayor según el "contraste de los multiplicadores de Lagrange" (índices de modificación) que informan de las saturaciones sobre otros factores diferentes a los especificados en el modelo.

Por último se pone a prueba el modelo de invarianza escalar entre los grupos, forzando la igualdad de los términos constantes de las ecuaciones de medición, denominados interceptos, en el marco del AFC con estructura de medias y covarianzas. De manera similar al paso anterior, en primer lugar se fuerzan a ser iguales entre los dos

grupos a los interceptos de todos los ítems (excepto los liberados en el paso anterior) y, en caso de no equivalencia, se dejan libres, uno a uno, los ítems cuyo desajuste sea mayor.

### 2.4.4. INVARIANZA MEDIANTE COMPARACIÓN DE MODELOS CON LA TRI

Se utiliza el procedimiento de Thissen *et al*., (1986) y Thissen *et al*., (1988, 1993) que comparan dos modelos con el estadístico de razón de verosimilitud: un *modelo con restricciones* en el que los parámetros de los ítems son idénticos en los grupos a comparar y un *modelo base* en el que los parámetros de los ítems del test pueden diferir a través de los grupos. La verosimilitud de los modelos se ha estimado con el programa MULTILOG (Thissen, 1991).

El estadístico de razón de verosimilitud $G^2$ es la diferencia entre los valores de verosimilitud de ambos modelos. Bajo la hipótesis nula, este estadístico sigue una distribución $\chi^2$ con grados de libertad igual a la diferencia entre el número de parámetros estimados en el modelo base y el número de parámetros estimados en el modelo con restricciones. Si el valor obtenido es mayor que el valor teórico de la distribución $\chi^2$ se rechaza la hipótesis nula, interpretándose que existe funcionamiento diferencial.

Con este procedimiento se prueban dos tipos de equivalencia de medida completa para comprobar si había diferencias con respecto a los parámetros *a* y *b*. El primer modelo de equivalencia completa obliga a la igualdad del parámetro *a* de todos los ítems entre los dos grupos, dejando libre el parámetro *b*; se utiliza, por tanto, para examinar si los parámetros de discriminación son invariantes entre los grupos.

Es posible que los parámetros *a* de todos los ítems sean invariantes, no habiendo funcionamiento diferencial no uniforme, pero que siga existiendo DIF debido al parámetro *b*. En ese caso, los sujetos de los distintos grupos tendrán la misma probabilidad de dar una determinada respuesta aún cuando tengan distinto nivel del rasgo latente, siendo siempre (para cualquier valor de $\theta$) el mismo grupo el que necesite un mayor nivel de rasgo latente para obtener la misma probabilidad de dar una respuesta determinada. Habría, por tanto, funcionamiento diferencial uniforme.

Por tanto, esta prueba tiene la ventaja de la TRI de detectar si la fuente del DIF se atribuye a diferencias en el parámetro *a* o en el parámetro *b*. Además, dado que el parámetro *a* es análogo al λ, esto es a la carga factorial del AFC, el modelo de invarianza completa para *a* aborda, desde otra perspectiva, el mismo objetivo que el modelo de equivalencia métrica multigrupo utilizado en el AFC. Asimismo, el modelo de invarianza completa de ambos parámetros *a* y *b* persigue los mismo objetivos que el modelo de equivalencia escalar del AFC multigrupo.

Por este motivo, se pone a prueba en primer lugar la equivalencia completa para el parámetro de *a*. Si se rechaza ésta entre los dos grupos, se procede a comprobar la invarianza de medida ítem a ítem, para localizar los ítems que provocan esta falta de equivalencia utilizando el programa IRTLRDIF, que utiliza el estadístico razón de verosimilitud para localizar qué ítems presentan DIF. De este modo, se libera la restricción de igualdad del parámetro *a* únicamente para el ítem que presente mayor DIF y se compara nuevamente con el modelo base en el marco de la equivalencia parcial de medida, continuando el proceso hasta que no haya diferencias significativas entre el modelo base y el modelo con restricciones de igualdad del parámetro entre ambos grupos. Dada la

naturaleza iterativa de este proceso, se ha ajustado el nivel de significación a .01 para controlar el error de tipo I (Cheung y Rensvold, 1999).

Una vez puesta a prueba la equivalencia del parámetro *a* entre ambos grupos, se procede del mismo modo, a comprobar la equivalencia para ambos parámetros *a* y *b*, primero comprobando la equivalencia completa y, de rechazarse ésta, en el ámbito de la equivalencia parcial. El procedimiento es exactamente el mismo que el ya explicado para el parámetro *a*, con la salvedad de que, aquí, el modelo de equivalencia completa excluye de restricción de igualdad de parámetros a los ítems que han resultado presentar DIF no uniforme. El motivo es que se ha detectado el DIF mediante una estructura jerárquica (utilizando IRTLRDIF), de manera que se comprueba la igualdad del parámetro *b* para comprobar el DIF uniforme solo si la diferencia del parámetro *a* entre los grupos no es significativa. Los tests sobre los parámetros b se ejecutan forzando la igualdad de los parámetros *a*; en este contexto, en caso de diferir el parámetro *a*, los posteriores análisis sobre el DIF del parámetro *b* no están garantizados (Teresi *et al*., 2007).

Como información gráfica complementaria se representa gráficamente la Curva Característica del Test (CCT) de cada grupo, para comprobar si difieren las puntuaciones esperadas en el test de ambos grupos a lo largo del continuo de impulsividad.

### 2.4.5. INVARIANZA MEDIANTE EL PROCEDIMIENTO DFIT

La estimación de los coeficientes de igualación se basó en el método de la Curva Característica del Test de Baker, implementado en el programa EQUATE 2.1 (Baker, 1995). Este programa dispone del procedimiento desarrollado por Stocking y Lord (1983)

para igualar la métrica de los parámetros, proporcionando dos coeficientes de transformación -pendiente y ordenada en el origen- para transformar linealmente los parámetros de un grupo en los de otro (para una descripción más detallada del proceso consultar Baker, 1992)). De esta forma se igualan todas las estimaciones de los parámetros del grupo focal a la métrica subyacente del grupo de referencia.

Varias investigaciones muestran que un procedimiento de igualación iterativo mejora la identificación de los ítems con DIF (e.g. Candell y Drasgow, 1988; Drasgow, 1987; Lautenschlager y Park 1988; Lord, 1980; Miller y Oshima, 1992). Por este motivo, para minimizar el error introducido por el procedimiento de igualación se utilizó un procedimiento de igualación en dos pasos. Después de la igualación inicial con todos los ítems del test se realizó un análisis del funcionamiento diferencial. Si resulta necesario eliminar algún ítem para determinar la equivalencia de medida, este ítem se elimina antes de ejecutar de nuevo el procedimiento de igualación, y de volver a calcular el funcionamiento diferencial de todos los ítems.

En un primer momento se computaron los estadísticos de funcionamiento diferencial de los ítems y del test mediante el procedimiento paramétrico DFIT de *et al*., (1995) con el programa DFITP5.

El estadístico NCDIF evalúa el DIF no compensatorio entre todos los ítems. Flowers *et al.* (1999) recomiendan utilizar un punto de corte de .054 en el índice NCDIF para ítems con cuatro opciones de respuesta. Además de este criterio, varios estudios de simulación (con el método Monte Carlo) sugieren que el valor de NCDIF debe ir acompañado por un valor significativo de $\chi^2$, $p$<.01. El estadístico DTF evalúa el

funcionamiento diferencial del test basándose en el índice compensatorio del ítem CDIF, que tiene en cuenta el posible funcionamiento diferencial de los ítems en direcciones opuestas. El punto de corte para el índice DTF es igual al indicado para el NCDIF multiplicado por el número de ítems de la escala. Si el valor de DTF indica funcionamiento diferencial del test se elimina el ítem con un valor mayor de CDIF, realizándose de nuevo el análisis de DTF. Este procedimiento iterativo continúa hasta que DTF deja de ser significativo.

Los puntos de corte de NCDIF predeterminados por Raju en base a estudios de simulación han resultado ser poco sensibles en la detección de ítems con DIF, con muchos falsos negativos, por lo que varios autores los consideran demasiado simplistas (Meade, Lautenschlager y Johnson, 2007; Oshima y Morris, 2008). Estudios posteriores demuestran que el punto de corte apropiado depende de factores tales como el tamaño de la muestra y el modelo de la TRI utilizado (Bolt, 2002; Chamblee, 1998).

Por este motivo, Raju, en colaboración con Oshima y Nanda (Oshima et al., 2006), proponen, para el caso dicotómico, el método de replicación de parámetros del ítem (Item Parameter Replication, IPR) que proporciona un medio de obtener valores de corte que se adaptan a un conjunto de datos particular. El método IPR se implementó en el año 2005 en la penúltima versión del software DFIT para ítems dicotómicos, y en el año 2009 en la última versión disponible para casos politómicos, llamada DFIT8 (Oshima, Kushubar, Scott y Raju, 2009).

Debido a las diferencias notables en la detección del funcionamiento diferencial de ítems y tests de las distintas versiones del software se decidió volver a analizar todos los

datos utilizando la versión DFIT8, capaz de obtener valores de corte para cada conjunto de datos en ítems politómicos. Para cada variable y escala (o subescala) se representa la CCT, para obtener información gráfica sobre la puntuación esperada en función del nivel de impulsividad por grupos.

# 3. RESULTADOS

Se exponen a continuación los resultados obtenidos en la investigación, comenzando por las propiedades psicométricas del test utilizado, -validez de constructo, fiabilidad, calidad de sus ítems y ajuste al MRG de Samejima-, mostrando posteriormente el impacto en las variables sexo y edad, para presentar finalmente los resultados de analizar la equivalencia métrica en ambas variables con los tres procedimientos analizados: el AFC multigrupo, la comparación de modelos mediante el test de razón de verosimilitud y el procedimiento DFIT.

## 3.1. PROPIEDADES PSICOMÉTRICAS DEL TEST BIS

Se aportan evidencias de validez de constructo estudiando la dimensionalidad del instrumento BIS mediante AFC, valorando la replicabilidad del estudio con la estrategia de validación cruzada. Se evalúa la consistencia interna del test BIS, así como la discriminación de los ítems utilizando el modelo clásico y el MRG de Samejima, realizando la estimación de los parámetros de los ítems con la muestra de calibración y estimando el nivel de aptitud de cada sujeto con la muestra de validación. Se realizan consideraciones sobre la adecuación del ajuste del modelo utilizado a los datos.

### 3.1.1. EVIDENCIAS DE VALIDEZ DE CONSTRUCTO: ESTUDIO DE LA DIMENSIONALIDAD DEL TEST MEDIANTE AFC

En primer lugar se divide la muestra aleatoriamente en dos partes -muestra de calibración y muestra de validación-, para analizar posteriormente la validación cruzada de los datos. En un primer análisis se decide, mediante comparación de modelos, cuál es la estructura factorial más apropiada. Después se evalúa el ajuste del modelo, considerando las posibles mejoras del mismo hasta llegar a un modelo que nos sirva como base para probar la equivalencia en los diversos grupos. Para estos dos primeros análisis se utiliza exclusivamente la muestra de calibración, valorando la replicabilidad del estudio mediante un proceso de validación cruzada en el que se pone a prueba la equivalencia entre muestras de calibración y validación. En caso de ser el resultado de este proceso satisfactorio, ambas muestras se volverán a unir para realizar los estudios de equivalencia en edad y sexo.

#### *3.1.1.1. Comparación de modelos*

La muestra se distribuyó aleatoriamente en las mitades de calibración y validación utilizando la función *selección de casos* del programa SPSS. Una vez eliminados los sujetos con valores perdidos, las muestras de calibración y validación contenían 851 y 839 participantes, respectivamente.

El tamaño de las muestras de calibración y validación fue suficiente porque ambas muestras tienen más de 300 sujetos, la ratio del tamaño muestral en relación con los ítems

es mayor que 20:1, y el número de indicadores por factor es siempre mayor o igual a 5 (Bentler y Chou, 1987; Bollen, 1989; Marsh, Hau, Balla y Grayson, 1998; Tanaka, 1987).

Para comprobar que la distribución de las variables sexo y edad era semejante en las muestras de calibración y validación se utilizó el estadístico $\chi^2$, encontrando que no había diferencias en la distribución de hombres y mujeres ($\chi^2$ = 0.778, p = .378) y preadolescentes y adolescentes ($\chi^2$ = 0.599, p = .439) en función de la división de la muestra realizada.

Puesto que la evidencia es insuficiente para considerar a priori que la estructura del test es trifactorial (ver apartado 2.3) esta hipótesis se pondrá a prueba mediante AFC. Dado que algunos autores abogan por una estructura bifactorial (Fosatti, 2002; Recio, *et al.*, 2004) y en vista del gráfico de sedimentación de la Figura 3.1, cuyo autovalor del primer factor es más del doble del segundo (5.65 vs. 2.22), no se puede descartar una estructura de dos factores, e incluso unidimensional.

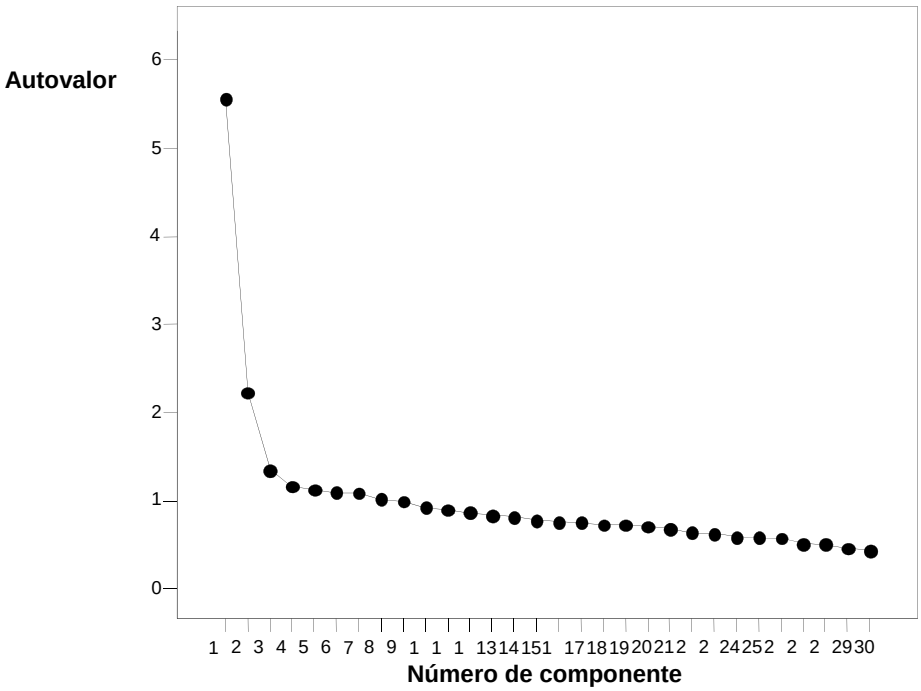



Fig. 3.1. Gráfico de sedimentación del test BIS.

En el modelo unifactorial todos los ítems saturan sobre un factor de impulsividad general. El modelo bifactorial distingue entre impulsividad motora y no motora, correspondiendo al primer factor los ítems 2, 6, 9, 12, 15, 18, 23, 26 y 29 y el resto al segundo. Por último, el modelo trifactorial considera que son tres los factores de impulsividad que se evalúan con el BIS, el impulso motor (ítems 2, 6, 9, 12, 15, 18, 23, 26 y 29), el impulso no planificado (ítems 1, 3, 5, 8, 11, 14, 17, 22, 25, 28, 30) y el impulso cognitivo-atencional (ítems 4, 7, 10, 13, 16, 19, 20, 21, 24 y 27).

El método de estimación utilizado en los tres casos fue el de mínimos cuadrados ponderados robusto (DWLS), basado en la matriz de correlaciones policórica y su matriz de covarianzas asintóticas, que es el procedimiento recomendado en variables ordinales

-como es el caso de los ítems del BIS-, ya que proporciona estimaciones correctas de los errores en muestras grandes (Jöreskog, 2002).

La matriz de correlaciones obtenida se sustenta en el supuesto de que existe normalidad bivariada subyacente entre todas las variables. Para evaluar esta normalidad se suele utilizar el estadístico de razón de verosimilitud. El problema de este estadístico es que tiende a rechazar la normalidad en muestras grandes, por lo que Joreskog (2002) ha desarrollado una forma de comprobar la normalidad subyacente basándose en un estadístico RMSEA similar al formulado por Steiger (1990). Según Joreskog (2002) hay efectos de no-normalidad si el valor de RMSEA es mayor que .1.

Se comprobó en los valores del estadístico RMSEA para cada par de ítems que ninguno superaba el valor de .1, por lo que se consideró que existe normalidad bivariada y que se puede utilizar la matriz de correlaciones policórica en el AFC.

Se ha fijado a 1 la saturación factorial de un ítem por cada variable latente para identificar su escala; siguiendo las recomendaciones de Byrne (1998) y Jöreskog y Sörbom (1996), estos ítems se eligieron teniendo en cuenta su alta fiabilidad, y se han señalizado en gris en los diagramas de vías correspondientes (ver Figuras 3.2, 3.3 y 3.4).

Los índices de ajuste global evaluados arrojan resultados apropiados en los tres modelos propuestos (ver Tabla 3.1.), con algunas diferencias. En cuanto a la estimación de los parámetros, en las Figuras 3.2, 3.3 y 3.4 se muestra el diagrama de vías de los tres modelos propuestos, incluyendo los coeficientes de regresión estandarizados y su error

estándar, así como la correlación estimada entre los factores en los casos bifactorial y trifactorial.

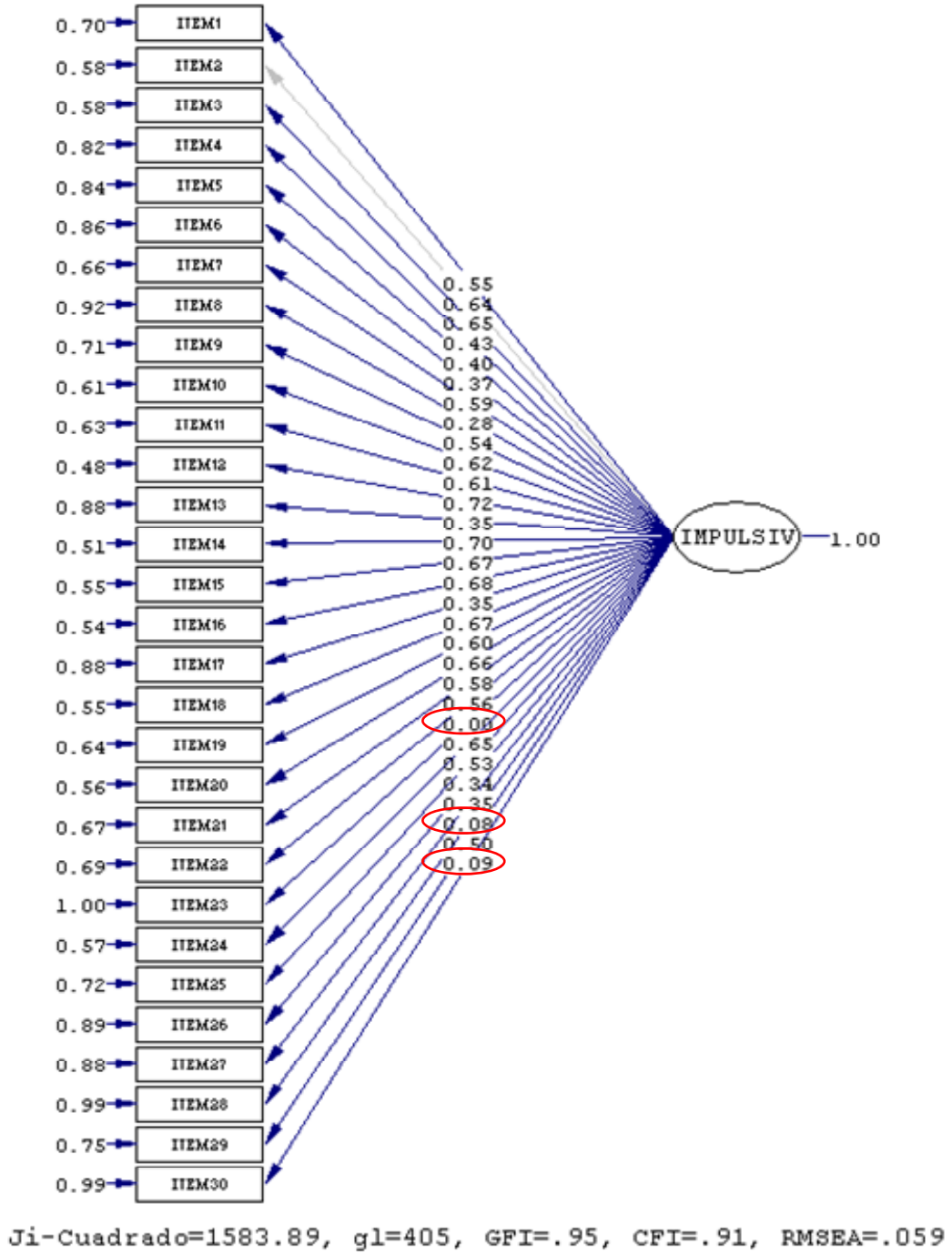



Figura 3.2. Diagrama de vías de la solución unifactorial del test BIS.

En el modelo unifactorial la saturación factorial del ítem 2 se fijó a 1 para identificar la escala de la variable latente. La revisión de los parámetros estimados revela que no

existen estimaciones fuera de rango; utilizando el estadístico de contraste *t*, con un nivel de confianza de .95 para comprobar si el valor de los coeficientes de regresión estandarizados son significativamente diferentes de cero, todos resultaron ser estadísticamente significativos, exceptuando el correspondiente al ítem 23 (ver Figura 3.2), aunque los ítems 28 y 30 también presentan valores muy bajos de saturación factorial (.10 y .07 respectivamente). Los 27 ítems restantes tienen valores apropiados de carga factorial, que oscilan entre .33 y .80.

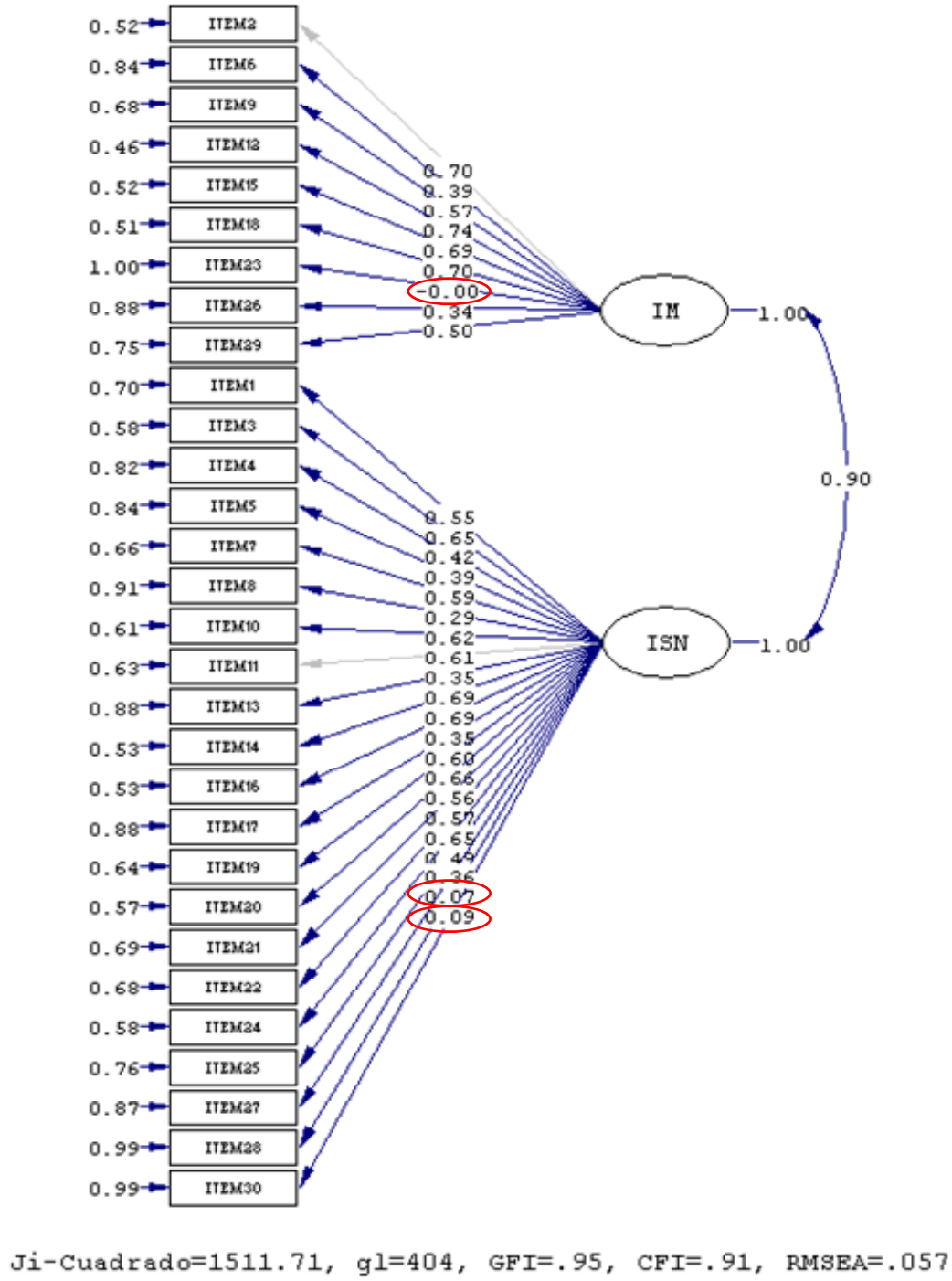



Figura 3.3. Diagrama de vías de la solución bifactorial del test BIS.

Los valores de los parámetros de la solución bifactorial también son adecuados con la misma excepción de los ítems 23, 28 y 30, con cargas factoriales de .00, .07 y .08, respectivamente que, en el primer caso no es significativamente distinta de cero. El resto

de valores oscilan entre .29 y .70. La correlación estimada entre ambos factores es muy elevada, lo que es coherente con la presencia de un gran factor común (ver Figura 3.3.)

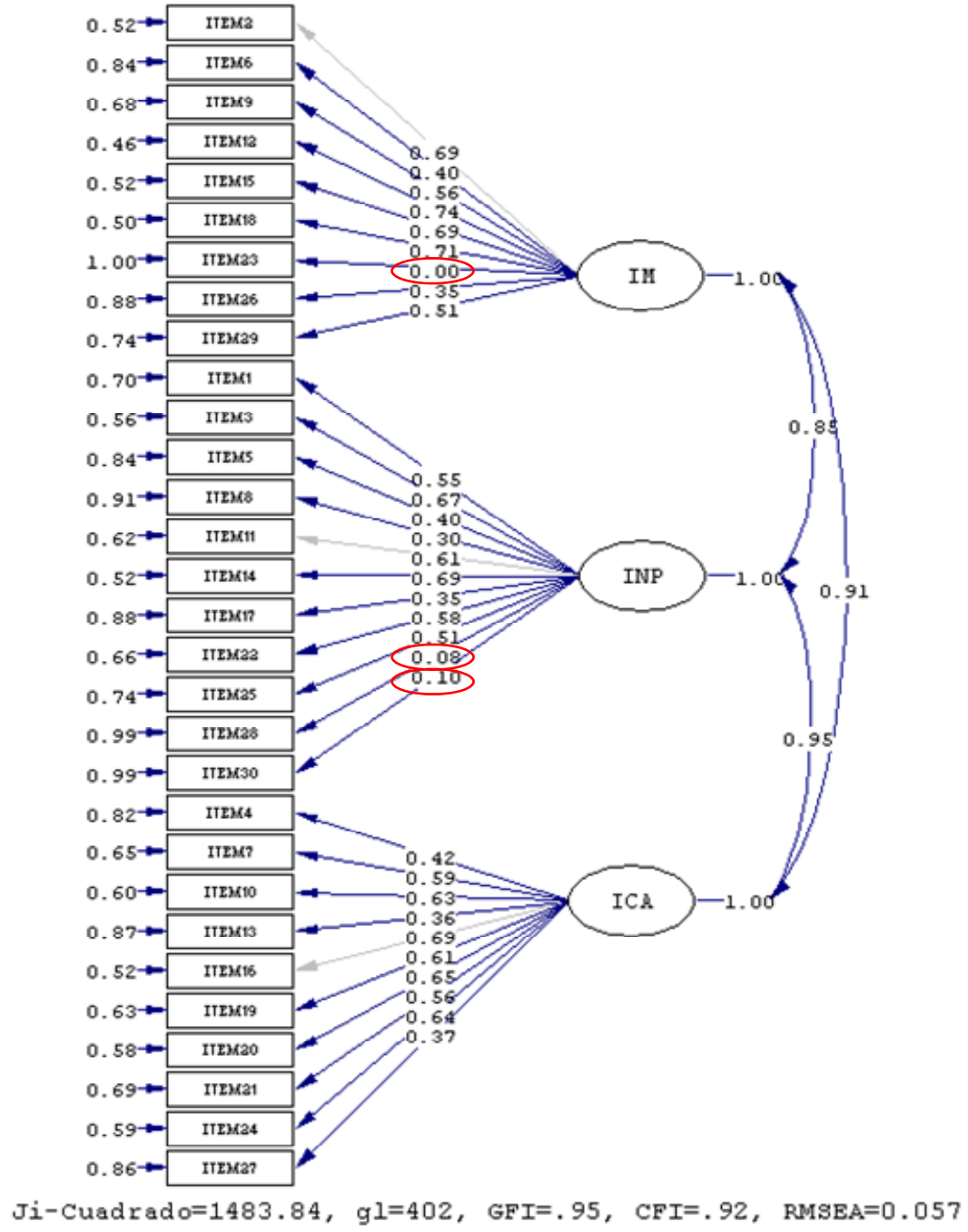



Figura 3.4. Solución trifactorial del test BIS.

Atendiendo a los valores estandarizados de los coeficientes de regresión del modelo trifactorial (ver Figura 3.4) se puede apreciar que sus valores son apropiados, ya que las

saturaciones factoriales son en general altas aunque, al igual que los casos anteriores, hay tres excepciones. Éstas se encuentran en los ítems 23, 28 y 30, que presentan valores extremadamente bajos de saturación factorial, esto es, .00, .08 y .10 respectivamente, en comparación con los del resto de los ítems de la escala con valores situados entre.30 y .74. En cuanto a las estimaciones de las correlaciones entre las subescalas Impulso Motor (IM), Impulso No Planificado (INP) e Impulso Cognitivo-Atencional (ICA) adoptan valores factibles, ya que ninguna es negativa o excede la unidad (ver Figura 3.4), aunque muy elevados.

En la Tabla 3.1 se muestran los valores de los índices de bondad de ajuste para los tres modelos propuestos. El modelo bifactorial se ajusta mejor a los datos que el modelo unifactorial, y el modelo trifactorial, a su vez, tiene un mejor ajuste que los dos anteriores.

Tabla 3.1. *Comparativa de los índices de ajuste de las tres estructuras factoriales propuestas para el BIS-PA*

| | $\chi^2$ | g.l. | $\Delta\chi^2$ | $\Delta$ g.l. | p | GFI | NNFI | CFI | ECVI | RMSEA |
|---|---|---|---|---|---|---|---|---|---|---|
| Est. unifactorial | 1583.89 | 405 | | | | .95 | .90 | .91 | 2.04 | .059 |
| Est. bifactorial | 1511.71 | 404 | 72.18 | 1 | .01 | .95 | .91 | .92 | 1.96 | .057 |
| Est. trifactorial | 1483.84 | 402 | 27.87 | 2 | .01 | .95 | .91 | .92 | 1.93 | .057 |

Atendiendo a la comparativa entre los modelos, hay diferencias estadísticamente significativas en el incremento de $\chi^2$ entre los tres modelos propuestos, obteniendo el modelo trifactorial el mejor ajuste. No obstante, tal y como apuntan Coenders *et al.* (2005), resulta una contradicción evitar utilizar $\chi^2$ para evaluar el ajuste del modelo en muestras grandes y en cambio sí utilizarlo en la comparación de modelos. Cada vez hay mayor acuerdo en basarse en el $\Delta$CFI para valorar si el ajuste del modelo es significativamente

mejor, considerando que esto sucede con incrementos de CFI superiores a .01 (Chen, 2007; Cheung y Rensvold, 2002; Meade, Johnson y Braddy, 2008) Teniendo en cuenta estas consideraciones, se observa en la Tabla 3.1 que el incremento en CFI no es, en ningún caso, superior a .01, por lo que, según este indicador, no habría una mejora sustancial al aumentar el número de factores de la escala y el test podría considerarse unidimensional.

Otro índice frecuentemente utilizado en la comparación de modelos es el índice de validación cruzada (ECVI), que mide las discrepancias entre la matriz de covarianzas de la muestra analizada y la que se obtendría en otra muestra de tamaño equivalente; se considera que el modelo con un valor más pequeño de ECVI tiene el mejor potencial para la replicación (Browne y Cudeck, 1989). Según este indicador, el modelo trifactorial es el que tiene una mayor replicabilidad, al presentar el menor valor (1.93 frente a 2.04 y 1.96), aunque estos mismos autores advierten que los índices de validación cruzada no deben utilizarse de forma rígida en un procedimiento de decisión que automáticamente elige el modelo que presenta el índice más bajo, siendo de vital importancia tener en cuenta otras consideraciones, tales como la plausibilidad del modelo.

Hay, por tanto, tres estructuras factoriales que, considerándose por separado, presentan un ajuste a los datos puede considerarse apropiado, a la luz de los índices globales calculados (GFI, NNFI, CFI, por encima de .90 y RMSEA de aproximadamente .05). Los índices utilizados para comparar los tres modelos propuestos no muestran una superioridad manifiesta de los modelos bidimensional y tridimensional, por lo que el test BIS podría considerase unidimensional. Además, las correlaciones entre los factores de los modelos bidimensional y tridimensional son muy elevadas, lo que indica la presencia de un

gran factor común. En resumen, parece que la estructura trifactorial es la que muestra un mejor ajuste, pero no está claro hasta qué punto esta diferencia es importante.

Dado que esta es una cuestión de validez, es de vital importancia considerar el punto de vista sustantivo a la hora de tomar decisiones sobre la dimensionalidad de la escala. En este sentido, los autores del test original, así como un buen número de adaptaciones del test a otros idiomas (Bayle *et al.*, 2000; Fossatti *et al.*, 2001, 2002; Someya *et al.*, 2001) abogan por una estructura trifactorial,. Se considerarán por tanto, en los análisis de AFC los tres factores de la escala BIS, teniendo en cuenta en los análisis realizados desde la TRI cada factor o subescala por separado, así como la escala completa.

En la evaluación de los tres modelos resulta llamativo la baja saturación factorial de tres ítems: el ítem 23, el ítem 28 y el ítem 30. El motivo de su desajuste podría deberse a una baja correlación ítem-test, o a estar cargando en un factor inapropiado. Dado que su correlación ítem-test es $r_{23X}$ =.055, $r_{28X}$ =.050 y $r_{30X}$ =.012 respectivamente y teniendo en cuenta, además, que ninguno de ellos presenta una correlación con cualquier otro ítem de la escala superior a 0.15 entendemos que su mal funcionamiento es debido a una baja correlación ítem-test y sería similar en cualquiera de las subescalas. Por tanto, se considera apropiado eliminar estos tres ítems del BIS, por lo que se excluirán de los sucesivos análisis realizados.

En la Figura 3.5 se presenta el diagrama de vías del test BIS una vez eliminados estos ítems, incluyendo los coeficientes de regresión estandarizados y su error estándar, la correlación estimada entre los factores y algunos índices de ajuste global del modelo.

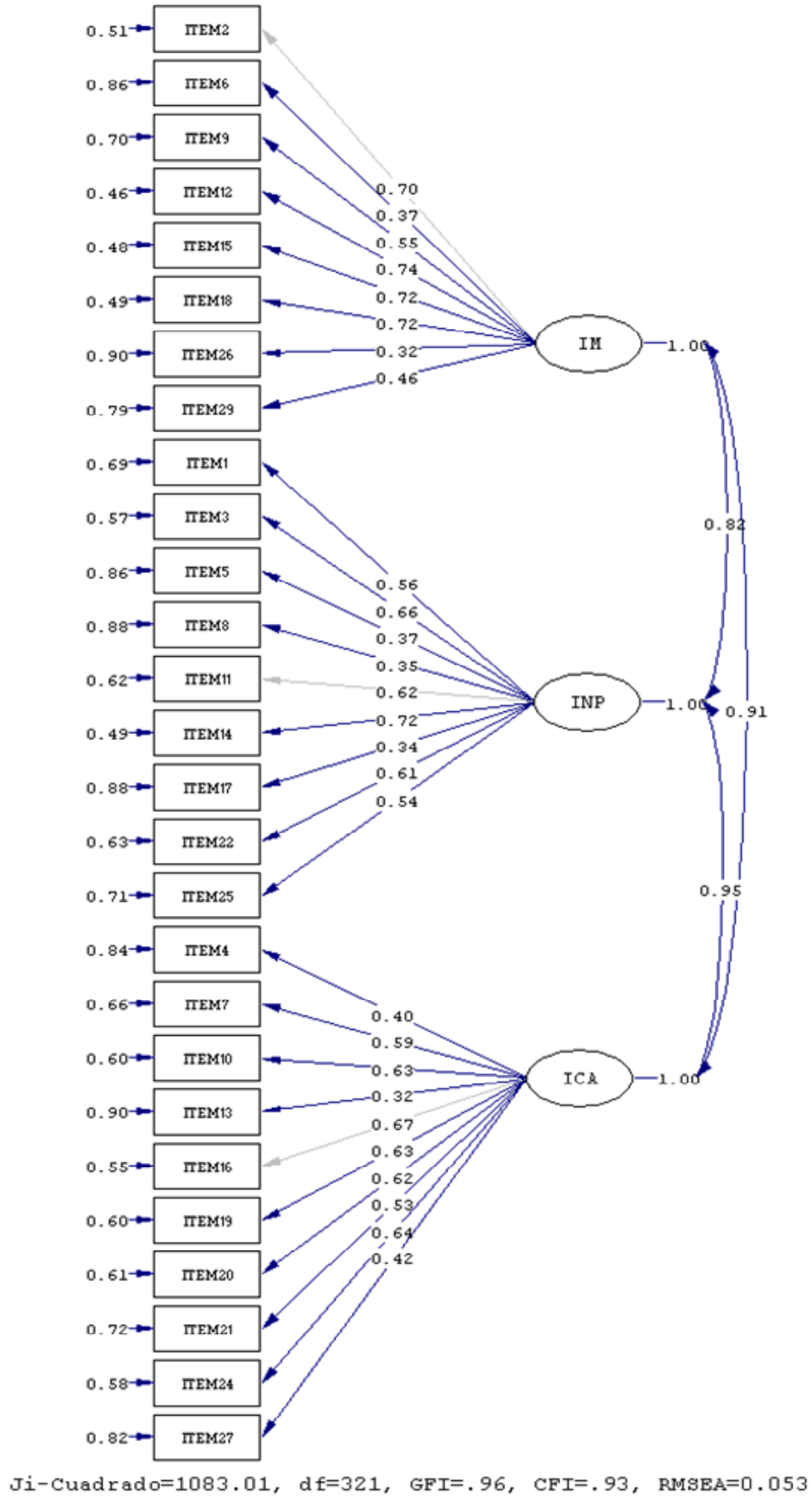



Figura 3.5. Solución trifactorial del test BIS una vez eliminados los ítems 23, 28 y 30.

Como era de esperar, tanto la estimación de los parámetros, con valores entre .32 y .74, como los índices globales de ajuste (GFI y CFI superiores a .90 y RMSEA inferior a .08) indican el buen ajuste del modelo a los datos, por lo que este será el modelo que sirva de base a los posteriores análisis factoriales confirmatorios.

#### *3.1.1.2. Validación cruzada*

Una vez realizados los análisis pertinentes con la muestra de calibración se comprueba la validez de los resultados utilizando la muestra de validación. Esta es la estrategia más común para valorar la generalizabilidad de un modelo. En este análisis multimuestra se fuerza, sucesivamente, la igualdad en la muestra de calibración y validación de todas las cargas factoriales (modelo 2), además de la matriz de covarianzas factoriales *phi* (modelo 6), y además de la matriz de covarianzas error *theta-delta* (modelo 4). El modelo base del que se parte considera la invarianza de configuración, es decir se fuerza la igualdad del número de factores y el patrón de matrices factoriales (modelo 1).

Tabla 3.2. *Índices de ajuste y comparación de modelos en la validación cruzada*

| | $\chi^2$ | g.l. | $\Delta\chi^2$ | $\Delta$ g.l. | p | GFI | NNFI | CFI | ECVI | RMSEA |
|---|---|---|---|---|---|---|---|---|---|---|
| Modelo 1 | 789.43 | 642 | | | | .99 | .96 | .96 | .60 | .016 |
| Modelo 2 | 833.67 | 666 | 44.24 | 24 | n.s. | .99 | .96 | .96 | .60 | .017 |
| Modelo 6 | 840.51 | 672 | 51.08 | 30 | n.s. | .99 | .96 | .96 | .60 | .017 |
| Modelo 4 | 969.12 | 699 | 179.69 | 57 | .01 | .98 | .93 | .93 | .64 | .021 |

Nota: modelo 1 = modelo base; modelo 2 = igualdad de cargas factoriales; modelo 6 = igualdad de cargas factoriales y matriz de covarianzas entre los factores (phi); modelo 4 = igualdad de cargas factoriales, matriz de covarianzas entre los factores (phi) y matriz de covarianzas error (theta-delta).

Considerando los resultados de los cuatros modelos por separado, se puede apreciar que los índices de bondad de ajuste son apropiados en todos los modelos a evaluar (GFI,

NNFI, CFI, por encima de .90 y RMSEA por debajo de .05). Estos índices son prácticamente idénticos en los tres primeros modelos y ligeramente peor ajustados en el modelo 4, al que se puede considerar además con peor potencial para la replicación, dado su mayor valor de ECVI (.64 respecto a .60 del resto de modelos).

Atendiendo a la comparativa entre los modelos, no hay diferencias significativas en el incremento de $\chi^2$ ni en el de CFI entre los modelos 1, 2 y 3, por lo que se puede afirmar que un modelo en el que se fuerza la igualdad tanto de los factores de carga de los ítems, como de las covarianzas entre los factores ajusta igual de bien que un modelo en el que estos parámetros se estiman de forma libre.

El único caso en el que hay diferencias significativas es en el modelo 4, que es el nivel de equivalencia mayor puesto a prueba, ya que establece la igualdad de cargas factoriales, de la matriz de covarianzas entre los factores y de la matriz de covarianzas error. Este nivel de equivalencia puede no ser realista en su aplicación práctica porque la medida y los errores específicos inherentes en las unicidades se asume que son aleatorios entre las muestras (MacCallum, Roznowski, Mar y Reith, 1994). Por tanto, habitualmente la evidencia de validación cruzada se considera aceptable para una medida, a condición de que las cargas factoriales de los ítems y las covarianzas de los factores se restrinjan a ser iguales entre las muestras de calibración y validación (Conroy y Molt, 2003), tal y como sucede en el modelo 3.

### 3.1.2. ANÁLISIS DE ÍTEMS

### *3.1.2.1. Análisis clásicos*

La Tabla 3.3 muestra dos índices clásicos en el análisis del funcionamiento de ítems: la puntuación media y la correlación ítem-total corregida (considerándose como total la puntuación en la subescala o en el test), para cada uno de los ítems de la subescala IM.

Partiendo de que los ítems de cada subescala miden la misma variable psicológica, se deben encontrar correlaciones positivas entre cada ítem y todos los demás. Esta es la base para el cálculo de los coeficientes de correlación entre las puntuaciones dadas en cada ítem y la suma de puntuaciones en todos los demás. Se utilizará la correlación corregida, esto es, la correlación entre las puntuaciones dadas en el ítem y la suma de puntuaciones de todos los demás ítems excepto el analizado. En general, se consideran aceptables todos los índices de discriminación iguales o superiores a .2 (e.g. Thorndike, 1989).

Tabla 3.3. *Puntuación media, desviación típica, correlación ítem-test corregida y correlación ítem-subescala corregida para cada uno de los ítems de la subescala Impulsividad Motora del BIS*

| ÍTEMS | $\overline{X}$ | $S_x$ | $r_{ix_{IM}}$ | $r_{ix_T}$ |
|---|---|---|---|---|
| Ítem 2 | 1.92 | 0.74 | .49 | .52 |
| Ítem 6 | 1.72 | 1.06 | .14 | .18 |
| Ítem 9 | 2.14 | 1.09 | .35 | .34 |
| Ítem 12 | 1.99 | 0.83 | .54 | .54 |
| Ítem 15 | 1.97 | 0.95 | .49 | .50 |
| Ítem 18 | 1.86 | 0.81 | .51 | .54 |
| Ítem 26 | 1.84 | 0.94 | .22 | .16 |
| Ítem 29 | 1.89 | 0.94 | .27 | .25 |

Notas: $r_{ix_{IM}}$ = correlación corregida del ítem con el total de la subescala Impulso Motor;

$r_{ix_T}$ = correlación corregida del ítem con el total de la escala BIS-PA;

Teniendo en cuenta que la valoración de todos los ítems puede oscilar entre 1 (nunca o casi nunca) y 4 (siempre o casi siempre), la simple visualización de las puntuaciones medias de los ítems evidencia que todos los valores se sitúan por debajo de la media teórica (2.5), en concreto, el promedio de esas puntuaciones medias es igual a 1.92, siendo la desviación típica de éstas igual a 0.12. Ello pone de manifiesto que los ítems de la escala representan conductas que son, en general, poco frecuentes.

En cuanto a las correlaciones entre las puntuaciones de cada ítem de la subescala Impulso Motor con la puntuación obtenida en el total de la subescala y en el total del test, todos los valores de correlación obtenidos son significativamente distintos de cero pero, aparte de ello, los valores en sí evidencian una capacidad discriminativa de los ítems en general aceptable (Wilmut, 1975), con la excepción del ítem 6, con bajas correlaciones con la subescala y con el test y el ítem 26 que, aunque presentan una aceptable correlación ítem-subescala (.22), su correlación ítem test es baja (.16). La media de las correlaciones para todos los ítems de esta subescala es igual a .38, al igual que la media de las correlaciones ítem-test.

En la Tabla 3.4 aparecen la puntuación media de cada ítem, su desviación típica y las correlaciones ítem-subescala e ítem-test, para cada uno de los ítems de la subescala INP.

Tabla 3.4. *Puntuación media, desviación típica, correlación ítem-test corregida y correlación ítem-subescala corregida para cada uno de los ítems de la subescala Impulsividad No Planificada del BIS*

| ÍTEMS | $\overline{X}$ | $S_x$ | $r_{ix_{INP}}$ | $r_{ix_T}$ |
|---|---|---|---|---|
| Ítem 1 | 2.30 | .83 | .38 | .42 |
| Ítem 3 | 1.91 | .81 | .37 | .47 |
| Ítem 5 | 2.55 | .97 | .19 | .21 |
| Ítem 8 | 2.39 | 1.10 | .24 | .28 |
| Ítem 11 | 1.69 | .88 | .45 | .42 |
| Ítem 14 | 1.72 | .92 | .38 | .44 |
| Ítem 17 | 2.22 | 1.17 | .21 | .19 |
| Ítem 22 | 1.84 | .90 | .45 | .45 |
| Ítem 25 | 1.40 | .77 | .18 | .25 |

Notas: $r_{ix_{INP}}$ = correlación corregida del ítem con el total de la subescala Impulso No Planificado;
$r_{ix_T}$ = correlación corregida del ítem con el total de la escala BIS-PA;

En esta subescala, el promedio de las puntuaciones medias de los ítems es algo mayor que en la subescala IM (2.08 frente a 1.92), si bien la puntuación media de todos los ítems (a excepción del ítem 5) se sitúa por debajo de la media teórica.

La mayoría de las correlaciones entre las puntuaciones de cada ítem de la subescala de Impulso No Planificado con la puntuación obtenida en el total de la subescala presentan valores apropiados, habiendo dos ítems cuyo valor está por debajo de .20 (el ítem 5 y 25); el resto de correlaciones ítem-subescala oscilan entre .21 y .45, siendo la media de las correlaciones de todos los ítems igual a .33. Al igual que sucedía en la subescala IM, las correlaciones ítem-total BIS son similares, adoptando valores desde .19 hasta .47, siendo 0.36 su promedio.

En la Tabla 3.5 aparecen la puntuación media de cada ítem, su desviación típica, la correlación ítem-subescala y la correlación ítem-test de cada uno de los ítems de la subescala ICA.

Tabla 3.5. *Puntuación media, desviación típica, correlación ítem-test corregida y correlación ítem-subescala corregida para cada uno de los ítems de la subescala Impulsividad Cognitivo-Atencional del BIS*

| ÍTEMS | $\overline{X}$ | $S_x$ | $r_{ix_{ICA}}$ | $r_{ix_T}$ |
|---|---|---|---|---|
| Ítem 4 | 2.37 | .98 | .14 | .23 |
| Ítem 7 | 2.49 | .98 | .41 | .45 |
| Ítem 10 | 2.17 | .92 | .31 | .37 |
| Ítem 13 | 2.67 | 1.00 | .16 | .14 |
| Ítem 16 | 2.08 | .98 | .45 | .45 |
| Ítem 19 | 2.34 | .93 | .41 | .49 |
| Ítem 20 | 1.76 | .81 | .44 | .47 |
| Ítem 21 | 1.89 | .96 | .29 | .37 |
| Ítem 24 | 1.72 | .85 | .36 | .41 |
| Ítem 27 | 2.06 | .96 | .23 | .32 |

Notas: $r_{ix_{ICA}}$ = correlación del ítem con el total de la subescala Impulso Cognitivo-Atencional;
$r_{ix_T}$ = correlación del ítem con el total de la escala BIS-PA;

En la subescala Impulso Cognitivo-Atencional, el promedio de las puntuaciones medias de los ítems es 2.22, valor algo mayor que el de las otras dos subescalas del BIS (ver Tablas 3.3 y 3.4). A pesar de este aumento la mayoría de los ítems presentan valores medios cercanos a 2, lo que equivale a responder "algunas veces" al ítem.

Los valores de las correlaciones ítem-subescala indican que la discriminación de estos ítems es, en general, apropiada aunque con dos excepciones, los ítems 4 y 13. La media de las correlaciones para todos los ítems de esta subescala es .33. La correlación ítem-test arroja valores que oscilan entre .14 y .49, con una media de .37.

#### *3.1.2.2. Estimación de parámetros*

Conceptualmente, el modelo TRI más apropiado para el BIS es un modelo para categorías de respuestas ordenadas como el MRG de Samejima aquí utilizado (Stark *et al*., 2002). Dado que las respuestas a los ítems del BIS se valoran en una escala de frecuencia con cuatro opciones de respuesta, hay cuatro parámetros para cada ítem: un parámetro de discriminación (*a*), que refleja la pendiente de la función de respuesta de la categoría y tres parámetros de localización ($b_1$, $b_2$ y $b_3$) que reflejan la posición de las funciones de respuesta a la categoría a lo largo del eje de abcisas.

Cuando se analizan los datos utilizando la TRI es conveniente lleva a cabo una validación cruzada, dividiendo los datos en una muestra de calibración y una muestra de validación, tal y como se ha realizado también para el AFC. En este caso, la muestra de calibración se utiliza para estimar los parámetros de los ítems y la muestra de validación para evaluar su ajuste empírico. Se utilizaron las mismas muestras de calibración y validación del AFC que contienen 851 y 839 participantes, respectivamente.

#### 3.1.2.2.1. Subescala Impulso Motor

En general, los ítems presentan valores de discriminación adecuados, siendo el promedio de todos ellos igual a 1.22 (D.T. = 0.73). En la Tabla 3.6 se presentan los valores estimados y sus errores típicos bajo el MRG de Samejima para cada uno de los 8 ítems que conforman la subescala IM.

Tabla 3.6. *Parámetros del ítem estimados y errores típicos asociados en la muestra de calibración para la subescala Impulsividad Motora del BIS*

| Item | *a* | *b1* | *b2* | *b3* |
|---|---|---|---|---|
| 2 | 1.83 (0.10) | -0.72 (0.06) | 1.26 (0.08) | 2.71 (0.18) |
| 6 | 0.47 (0.08) | 1.05 (0.25) | 3.21 (0.59) | 4.30 (0.80) |
| 9 | 0.80 (0.08) | -0.88 (0.14) | 1.03 (0.14) | 2.25 (0.24) |
| 12 | 2.30 (0.12) | -0.65 (0.05) | 0.98 (0.06) | 2.09 (0.11) |
| 15 | 1.32 (0.09) | -0.44 (0.08) | 1.13 (0.10) | 2.28 (0.17) |
| 18 | 2.11 (0.13) | -0.40 (0.05) | 1.18 (0.07) | 2.18 (0.13) |
| 26 | 0.37 (0.07) | -0.61 (0.26) | 4.11 (0.87) | 6.61 (1.40) |
| 29 | 0.54 (0.07) | -0.67 (0.18) | 2.55 (0.37) | 4.55 (0.65) |

Nota: los errores típicos aparecen entre paréntesis.

Hay dos ítems que muestran una discriminación muy alta, por encima de 2: el ítem 12 ($a$ = 2.30) y el ítem 18 ($a$ = 2.11) (ver Figura 3.6). El ítem 26 es el que presenta un valor menor de discriminación ($a$ = 0.37), aunque es ligeramente mejor en comparación con el análisis realizado en la escala completa (ver apartado 3.1.2.2.4).

En cuanto a los valores del parámetro de localización se encontraron entre el valor mínimo de –0.88 del ítem 9 y el máximo de 6.61 del ítem 26, que necesita de valores muy altos de Impulso Motor para que los sujetos respondan a las categorías 3 “bastantes veces” y 4 “siempre o casi siempre”. En este ítem, aunque la tendencia se invierte en los valores más altos de $\theta$, la probabilidad de responder a las categorías 1 “nunca o casi nunca” o 2 “algunas veces” es mayor.

La representación gráfica de las CCR de los ítems de la subescala Impulso Motor se muestra en la Figura 3.6.

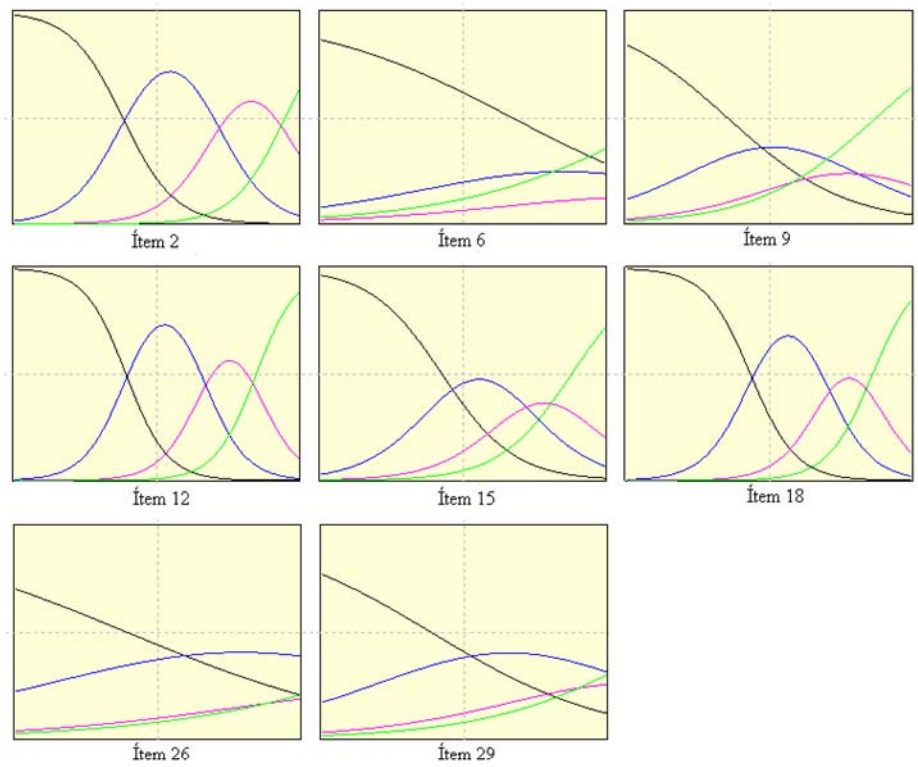



Figura 3.6. CCR de los 8 ítems que conforman la subescala Impulso Motor del BIS.

### 3.1.2.2.2. Subescala Impulso No Planificado

En la Tabla 3.7. se presentan los valores estimados y sus errores típicos bajo el MRG de Samejima para cada uno de los 9 ítems de la subescala Impulso No Planificado del BIS.

Tabla 3.7. *Parámetros del ítem estimados y errores típicos asociados en la muestra de calibración para la subescala Impulso No Planificado del BIS*

| Item | *a* | *b1* | *b2* | *b3* |
|---|---|---|---|---|
| 1 | 0.97 (0.09) | -1.48 (0.15) | 0.13 (0.09) | 4.14 (0.40) |
| 3 | 1.05 (0.09) | -0.86 (0.10) | 1.61 (0.14) | 3.25 (0.28) |
| 5 | 0.42 (0.07) | -3.87 (0.71) | -0.49 (0.21) | 3.54 (0.63) |
| 8 | 0.60 (0.07) | -1.74 (0.25) | 0.20 (0.14) | 2.36 (0.31) |
| 11 | 1.31 (0.10) | 0.15 (0.07) | 1.35 (0.11) | 2.81 (0.22) |
| 14 | 1.64 (0.10) | 0.14 (0.06) | 1.31 (0.09) | 2.17 (0.14) |
| 17 | 0.57 (0.07) | -0.57 (0.16) | 0.78 (0.17) | 2.86 (0.39) |
| 22 | 1.63 (0.10) | -0.11 (0.06) | 0.90 (0.07) | 2.53 (0.17) |
| 25 | 0.52 (0.09) | 2.02 (0.36) | 4.52 (0.79) | 6.05 (1.09) |

Nota: los errores típicos aparecen entre paréntesis.

El rango del parámetro de discriminación varía de 0.42 a 1.64 (ver Tabla 3.12.) siendo la media de todos ellos igual a 0.97 (D.T. = 0.45). La discriminación mayor la muestran los ítems 14 y 22, con valores de *a* de 1.64 y 1.63 respectivamente. El ítem 5 es el que presenta una peor discriminación ($a = 0.42$).

Los valores del parámetro *b* oscilan entre el valor –3.87 del ítem 5 y 6.05 del ítem 25. Estos dos ítems son los que presentan un menor valor de discriminación de la subescala ($a = 0.42$ y $a = 0.52$ respectivamente). En el caso del ítem 5 son necesarios unos niveles muy bajos de impulso no planificado para que el sujeto escoja la categoría 1 “nunca o casi nunca” y niveles bastante altos para escoger las categorías de respuesta 3 “bastantes veces” y 4 “siempre o casi siempre”. El ítem 25 necesita de niveles muy altos del rasgo para que los sujetos respondan a las categorías 3 y 4 del ítem.

En la Figura 3.7 se muestra la representación gráfica de las CCR de los ítems de la subescala Impulso No Planificado.

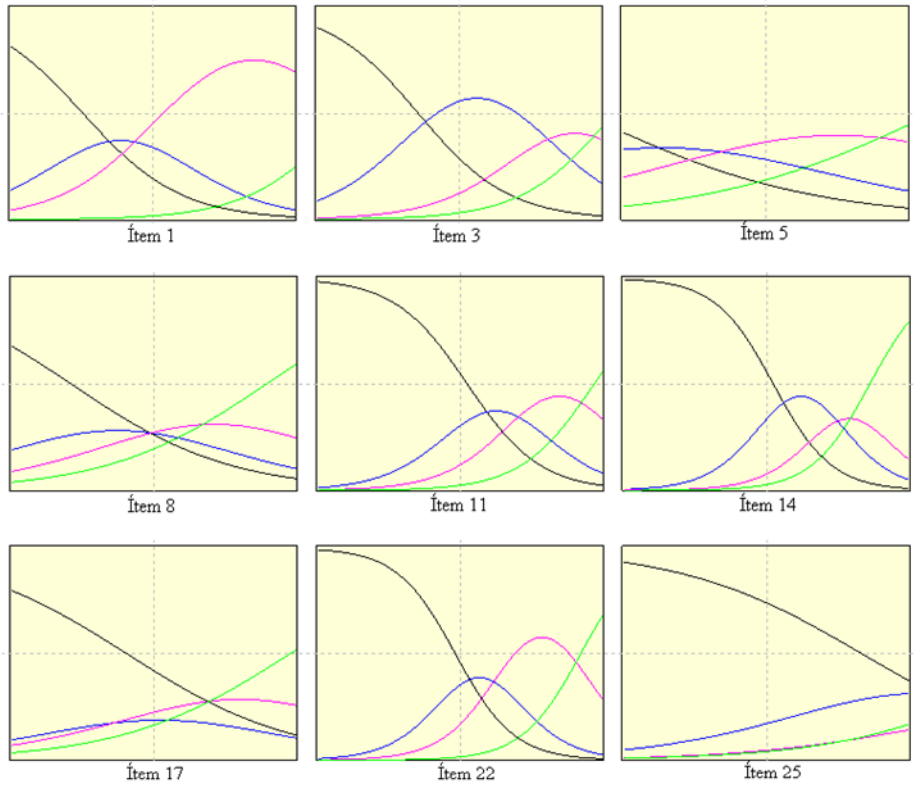



Figura 3.7. CCR de los 9 ítems que conforman la subescala Impulso No Planificado del BIS.

### 3.1.2.2.3. Subescala Impulso Cognitivo Atencional

En la siguiente tabla aparecen los valores estimados y sus errores típicos bajo el MRG de Samejima para los ítems de la subescala ICA.

Tabla 3.8. *Parámetros del ítem estimados y errores típicos asociados en la muestra de calibración para la subescala Impulso Cognitivo-Atencional del BIS*

| Item | *a* | *b1* | *b2* | *b3* |
|---|---|---|---|---|
| 4 | 0.34 (0.07) | -3.96 (0.81) | 0.64 (0.28) | 5.31 (1.16) |
| 7 | 1.34 (0.09) | -1.44 (0.11) | -0.24 (0.07) | 1.72 (0.12) |
| 10 | 1.05 (0.08) | -1.07 (0.12) | 0.58 (0.09) | 2.93 (0.24) |
| 13 | 0.58 (0.06) | -3.03 (0.40) | -0.84 (0.17) | 2.12 (0.29) |
| 16 | 1.35 (0.10) | -0.77 (0.08) | 0.97 (0.09) | 1.90 (0.13) |
| 19 | 1.08 (0.08) | -1.32 (0.12) | 0.19 (0.08) | 2.32 (0.19) |
| 20 | 1.25 (0.09) | -0.28 (0.07) | 1.83 (0.14) | 2.93 (0.22) |
| 21 | 0.60 (0.07) | -0.37 (0.14) | 2.10 (0.28) | 4.24 (0.55) |
| 24 | 0.98 (0.09) | -0.02 (0.09) | 1.86 (0.17) | 3.33 (0.31) |
| 27 | 0.48 (0.07) | -1.62 (0.29) | 1.99 (0.34) | 4.37 (0.68) |

Nota: los errores típicos aparecen entre paréntesis.

La media en discriminación (parámetro *a*) de los ítems de la subescala de Impulso cognitivo es igual a 0.91 (D.T. = 0.36). El peor indicador de Impulso Cognitivo-Atencional lo constituye el ítem 4 con un valor de discriminación de 0.34, pudiéndose apreciar en la Figura 3.8 que las curvas de probabilidad de las categorías están muy poco concentradas a lo largo de la escala $\theta$. En este ítem la probabilidad de escoger una opción de respuesta cualquiera no supera apenas el valor .4.

Como ejemplo de ítem apropiado para medir el Impulso Cognitivo-Atencional se puede destacar el ítem 7 (ver Figura 3.8). En este ítem los sujetos con un bajo nivel de rasgo tienen una probabilidad muy alta de responder a la categoría 1 del ítem (nunca o casi nunca), mientras que los sujetos con alto nivel de la característica de interés tienen probabilidades cercanas a cero de escogerla. Las categorías 2 y 3 son escogidas por sujetos con niveles intermedios de Impulso Cognitivo-Atencional mientras que las probabilidades mayores de elegir la categoría 4 del ítem están relacionadas con niveles altos de la característica evaluada.

En la Figura 3.8 se muestra la representación gráfica de las CCR de los ítems de la subescala Impulso Cognitivo-Atencional.

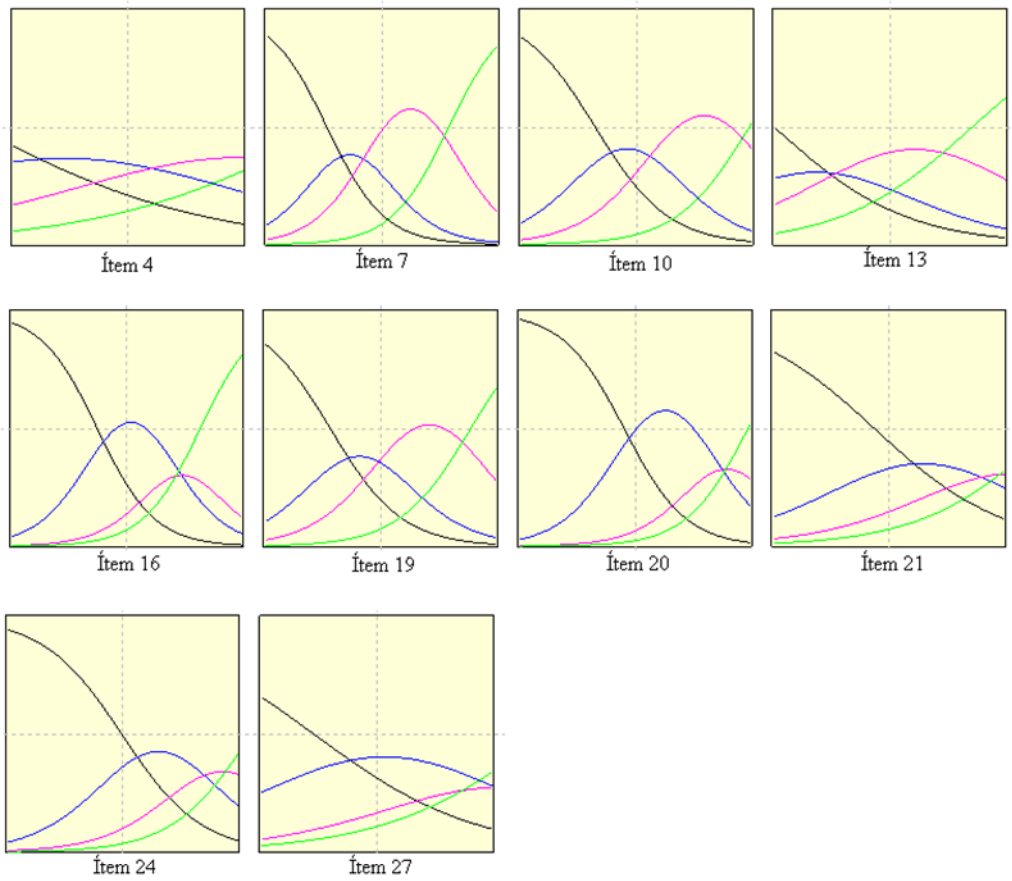



Figura 3.8. CCR de los 9 ítems que conforman la subescala Impulso Cognitivo-Atencional del BIS.

### 3.1.2.2.4. Escala BIS completa

Como se puede apreciar en la Tabla 3.9, el rango del parámetro *a* oscila entre 0.32 y 1.75. El ítem 26 es el que presenta una peor discriminación ($a$ = 0.32). Un ejemplo de ítem que presenta una excelente discriminación es el ítem 18 ($a$ = 1.75), ya que la probabilidad de marcar un 1 en el ítem cuando se tiene un bajo nivel de rasgo es muy alta, las opciones

2 y 3 del ítem necesitan niveles moderados a altos de impulsividad para ser escogidas y se necesita un alto nivel de impulsividad para marcar la opción 4 del ítem (ver Figura 3.9).

Tabla 3.9. *Parámetros del ítem estimados y errores típicos asociados en la muestra de calibración para el total de la escala BIS*

| Item | *a* | *b1* | *b2* | *b3* |
|---|---|---|---|---|
| 1 | 1.00 (0.09) | -1.50 (0.15) | 0.08 (0.09) | 4.02 (0.39) |
| 2 | 1.44 (0.10) | -0.85 (0.08) | 1.39 (0.11) | 3.07 (0.24) |
| 3 | 1.22 (0.09) | -0.83 (0.09) | 1.40 (0.12) | 2.87 (0.23) |
| 4 | 0.49 (0.08) | -2.86 (0.49) | 0.40 (0.19) | 3.70 (0.61) |
| 5 | 0.45 (0.08) | -3.64 (0.64) | -0.50 (0.20) | 3.25 (0.57) |
| 6 | 0.49 (0.09) | 0.96 (0.24) | 3.03 (0.56) | 4.07 (0.75) |
| 7 | 1.15 (0.09) | -1.63 (0.14) | -0.31 (0.08) | 1.85 (0.15) |
| 8 | 0.63 (0.08) | -1.71 (0.24) | 0.14 (0.14) | 2.20 (0.29) |
| 9 | 0.78 (0.08) | -0.93 (0.14) | 1.00 (0.15) | 2.24 (0.25) |
| 10 | 1.00 (0.09) | -1.14 (0.13) | 0.55 (0.10) | 2.99 (0.27) |
| 11 | 1.03 (0.09) | 0.13 (0.09) | 1.54 (0.15) | 3.29 (0.32) |
| 12 | 1.53 (0.11) | -0.82 (0.08) | 1.14 (0.09) | 2.49 (0.18) |
| 13 | 0.36 (0.07) | -4.72 (0.95) | -1.33 (0.34) | 3.23 (0.69) |
| 14 | 1.23 (0.10) | 0.12 (0.07) | 1.49 (0.13) | 2.54 (0.21) |
| 15 | 1.23 (0.09) | -0.51 (0.08) | 1.14 (0.11) | 2.36 (0.19) |
| 16 | 1.09 (0.09) | -0.93 (0.11) | 1.06 (0.11) | 2.13 (0.18) |
| 17 | 0.44 (0.08) | -0.75 (0.23) | 0.94 (0.25) | 3.56 (0.64) |
| 18 | 1.75 (0.11) | -0.46 (0.06) | 1.24 (0.08) | 2.31 (0.16) |
| 19 | 1.19 (0.09) | -1.26 (0.11) | 0.14 (0.08) | 2.12 (0.17) |
| 20 | 1.12 (0.09) | -0.34 (0.08) | 1.93 (0.16) | 3.11 (0.26) |
| 21 | 0.79 (0.08) | -0.34 (0.11) | 1.61 (0.19) | 3.28 (0.36) |
| 22 | 1.05 (0.09) | -0.19 (0.09) | 1.11 (0.12) | 3.34 (0.32) |
| 24 | 0.98 (0.09) | -0.06 (0.09) | 1.83 (0.18) | 3.30 (0.33) |
| 25 | 0.63 (0.10) | 1.66 (0.27) | 3.76 (0.58) | 5.05 (0.81) |
| 26 | 0.32 (0.16) | -0.76 (0.29) | 4.60 (1.77) | 7.43 (3.22) |
| 27 | 0.70 (0.08) | -1.20 (0.18) | 1.39 (0.19) | 3.07 (0.37) |
| 29 | 0.53 (0.08) | -0.72 (0.19) | 2.52 (0.38) | 4.55 (0.67) |

Nota: los errores típicos aparecen entre paréntesis.

En la Figura 3.9 se muestra la representación gráfica de las CCR de todos los ítems de la escala BIS analizados conjuntamente.

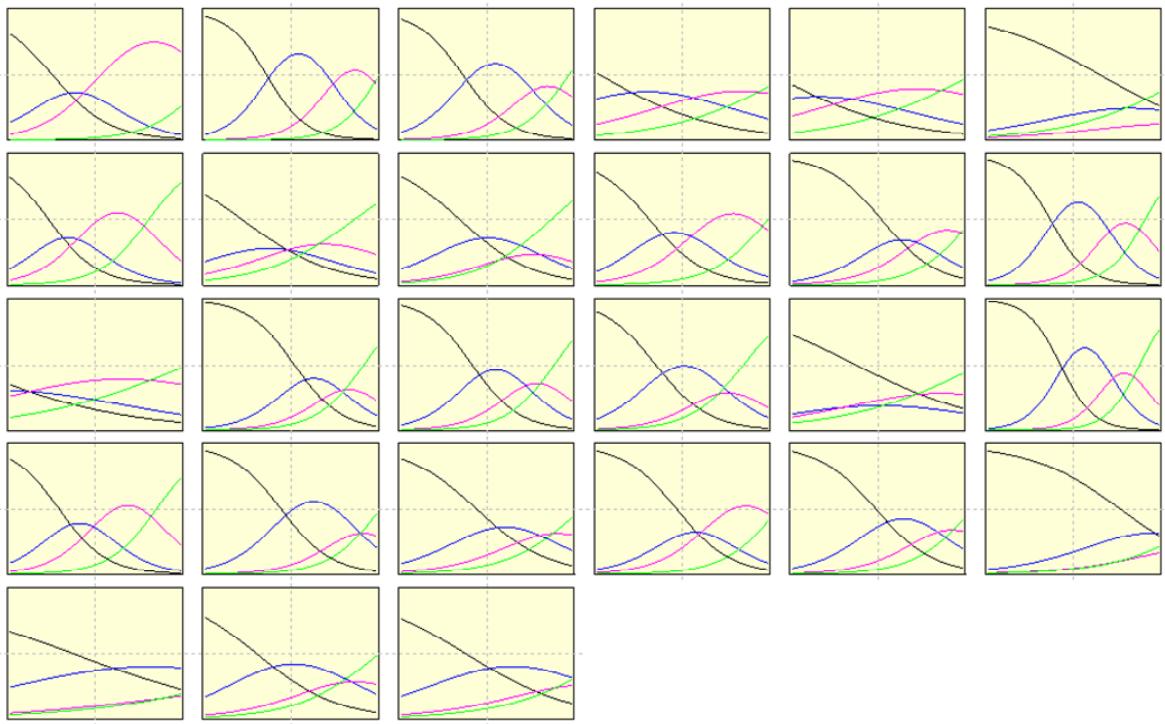


Figura 3.9. CCR de los 27 ítems de la escala BIS.

### 3.1.3. FIABILIDAD

#### *3.1.3.1. Coeficiente Alfa*

La fiabilidad de la escala en su conjunto es alta, habiéndose obtenido un coeficiente α igual a .83. Por subescalas encontramos que α = .67 en Impulsividad Motora, α = .63 en Impulsividad No Planificada y α = .65 en Impulsividad Cognitivo-Atencional. En la valoración de estos coeficientes se debe tener en cuenta los tamaños de estas escalas; así, la fiabilidad del test total es razonablemente más alta que la de las tres subescalas, lo que en parte podría explicarse por un mayor número de ítems del total de la escala (27 ítems) frente a los 8, 9 y 10 ítems que conforman las tres subescalas IM, INP e ICA respectivamente.

### *3.1.3.2. Procedimientos factoriales*

La fiabilidad se ha estimado por medio de distintos indicadores derivados de procedimientos factoriales: alfa ordinal, theta y omega (ver Tabla 3.10). Estos procedimientos tienen en cuenta la naturaleza ordinal de las variables, y sus cálculos se basan en la matriz de correlaciones policórica. Puede encontrarse una explicación detallada de los dos primeros procedimientos en Elosua y Zumbo (2008) y del último en Brown (2006) y Gómez (1996).

Tabla 3.10. *Fiabilidad de las subescalas del test BIS y de la escala completa*

| | alfa ordinal | theta | omega[1] |
|---|---|---|---|
| Impulso Motor | .74 | .77 | .80 |
| Impulso No Planificado | .71 | .73 | .81 |
| Impulso Cognitivo Atencional | .70 | .72 | .82 |
| Escala completa | .87 | .88 | .93 |

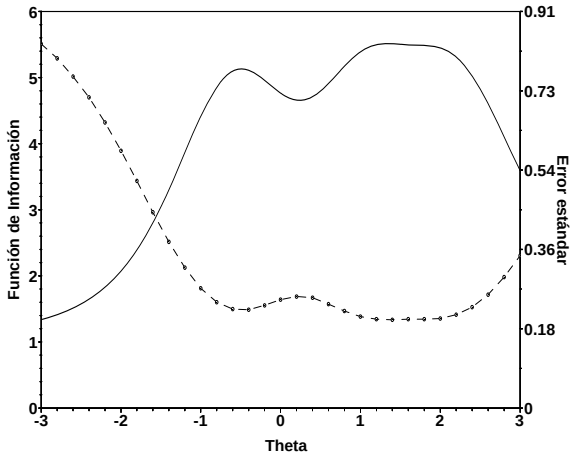
[1] Calculado con los resultados del AFC.

La estimación de la fiabilidad, tanto de las subescalas como de la escala completa es mayor con estos coeficientes que con el coeficiente alfa. Hay diferencias en los valores estimados según los métodos factoriales empleados, proporcionando los valores mayores de fiabilidad el procedimiento basado en el AFC (ver Tabla 3.10). Estos resultados están en línea con lo esperado (ver apartado 2.5.1.3.), teniendo en cuenta la subestimación de la fiabilidad en datos ordinales (Bentler, 2009; Zumbo *et al.*, 2007).

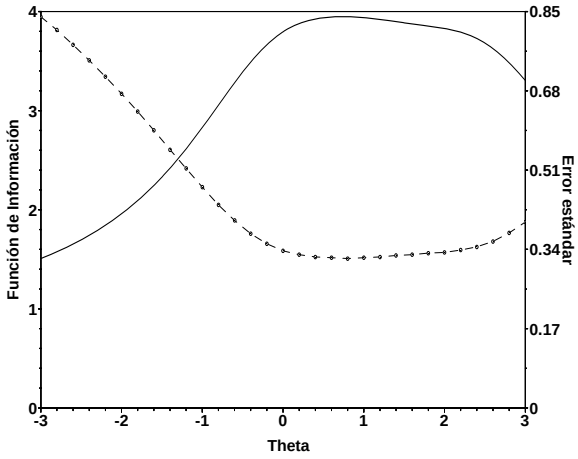
### *3.1.3.3. Función de información*

Se ha calculado la función de información de cada una de las subescalas del test BIS, así como de la escala en su conjunto (ver Figuras 3.10 y 3.11).

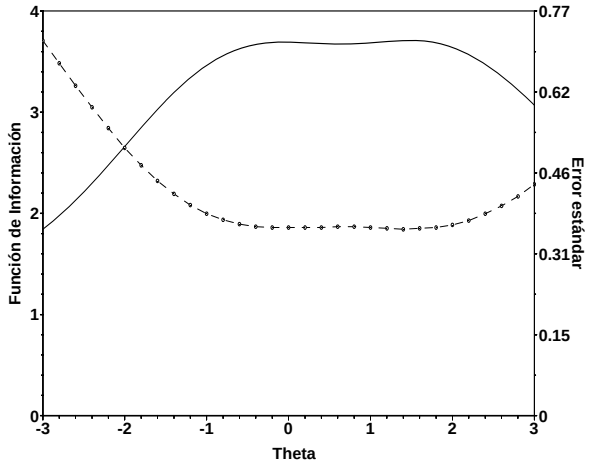
(a)




(b)




(c)




Nota: la información total de la subescala se lee en el eje vertical izquierdo (línea continua) y el error estándar en el eje vertical derecho (línea punteada).

Figura 3.10. Función de información total y error estándar de las tres subescalas del BIS: Impulso Motor (a), Impulso No Planificado (b) e Impulso Cognitivo-Atencional (c) en función del nivel de $\theta$.

La subescala Impulso Motor es la que proporciona mayores niveles de información. En general, las tres subescalas del BIS resultan más informativas en los niveles medios y altos del rasgo, si bien la subescala Impulso Cognitivo-Atencional es la que necesita de menores niveles de rasgo para llegar a su nivel de información máxima.

En cuanto a la función de información del test completo, la escala BIS produce una cantidad de información razonable entre –1 y +3, por lo que su uso es de amplia aplicabilidad, siendo únicamente desaconsejable en personas con niveles muy bajos del rasgo medido.

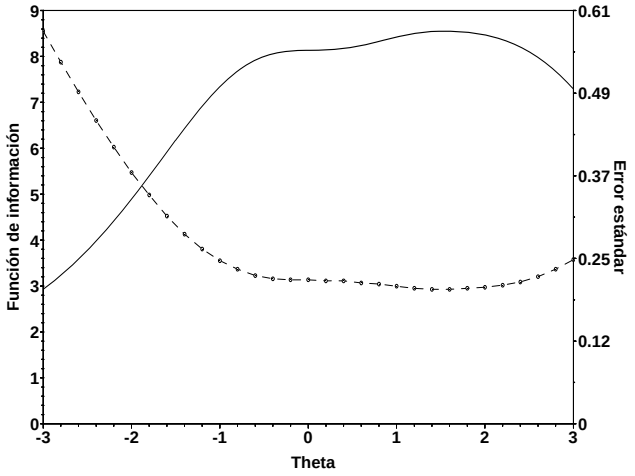



Figura 3.11. Función de información total y error estándar del test BIS completo, en función del nivel de actitud $\theta$.

### 3.1.4. AJUSTE DEL MODELO DE RESPUESTA GRADUADA DE SAMEJIMA A LOS DATOS

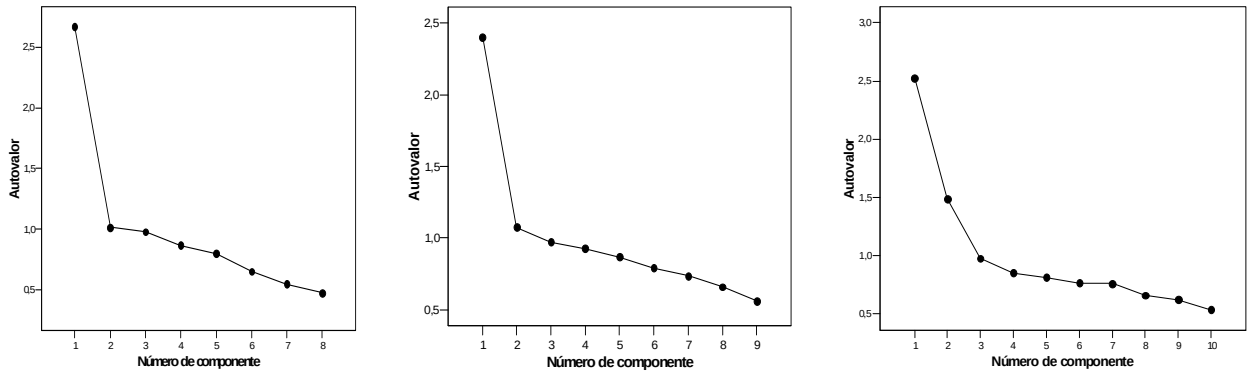
Para valorar si es apropiada la utilización del MRG, se ha comprobado el cumplimiento necesario del supuesto de unidimensionalidad, y se ha evaluado la

adecuación del ajuste a los datos atendiendo a información de tipo estadística (índices $\chi^2$ que evalúan el ajuste con respecto a las frecuencias conjuntas de las puntuaciones del ítem de primer orden, segundo orden y tercer orden, respectivamente) y de tipo gráfica (representando la correspondencia entre la curva teórica y la curva empírica de cada alternativa del ítem).

#### *3.1.4.1. Unidimensionalidad*

Por lo visto en el apartado 3.1.1, el test BIS puede ser visto como unidimensional, requisito necesario para realizar un análisis basado en la TRI. Además los indicios allí refereidos, se han considerado dos criterios adicionales: (1) la varianza explicada por el primer factor en un análisis de componentes principales debe ser mayor que el 20% (Drasgow y Parsons, 1983; Reckase, 1979) y (2) el gráfico de sedimentación de los autovalores debe reflejar un primer factor dominante (Hambleton, 1989).

Como era de esperar, no solo la escala completa sino las tres subescalas cumplen estos requisitos. Según el análisis factorial de componentes principales, en la subescala Impulso Motor el primer factor explica el 33.4%, en Impulso No Planificado el 26.7% y en Impulso Cognitivo-Atencional el 25.2%, valores todos ellos superiores al 20% requerido. En la Figura 3.12 se muestra el gráfico de sedimentación de los autovalores para las tres subescalas que componen el test.




**(a) Impulso Motor** **(b) Impulso No Planificado** **(c) Impulso Cognitivo-Atencional**

Figura 3.12. Gráfico de sedimentación de los autovalores para la subescala Impulso Motor (a), la subescala Impulso No Planificado (b), y la subescala Impulso Cognitivo-Atencional (c).

Tal y como se puede apreciar en la Figura 3.12, el primer factor está claramente distanciado del resto, por lo que se cumplen los dos requisitos propuestos, lo que posibilita utilizar el MRG de Samejima.

### *3.1.4.2. Valoración del ajuste*

#### 3.1.4.2.1. Subescala Impulso Motor del BIS

Se utilizan tres tipos de índices $\chi^2$ calculados con el programa MODFIT (Stark, 2001) para evaluar el ajuste del modelo. En el caso de los ítems individuales, la media de $\chi^2$/g.l. fue 1.473. Todos los ítems presentan un ajuste según este índice adecuado (<3) excepto el ítem 12, cuyo valor de $\chi^2$ fue 5.886. El desajuste estadístico de este ítem es muy elevado, por lo que si se omite su valor en la media de $\chi^2$ ésta tendría un valor de 0.921.

El ajuste gráfico de los ítems de la subescala de Impulso Motor es adecuado. Lógicamente, el ítem 12 presenta un ajuste gráfico peor que el resto de los ítems (ver Figura 3.12), si bien la correspondencia entre las CCR teóricas (ORF) y empíricas (EMP) es alta. En el caso de las categorías de respuesta 1 y 2 el ajuste es prácticamente perfecto, estando ambas líneas solapadas, mientras que hay un pequeño desajuste gráfico en las categorías que indican fuerte presencia del rasgo (3 y 4). En cualquier caso, como se verá en el apartado 3.1.4.2.4, en comparación con el ajuste gráfico de este ítem en la escala completa (ver Figura 3.17) este desajuste gráfico es bastante menor.

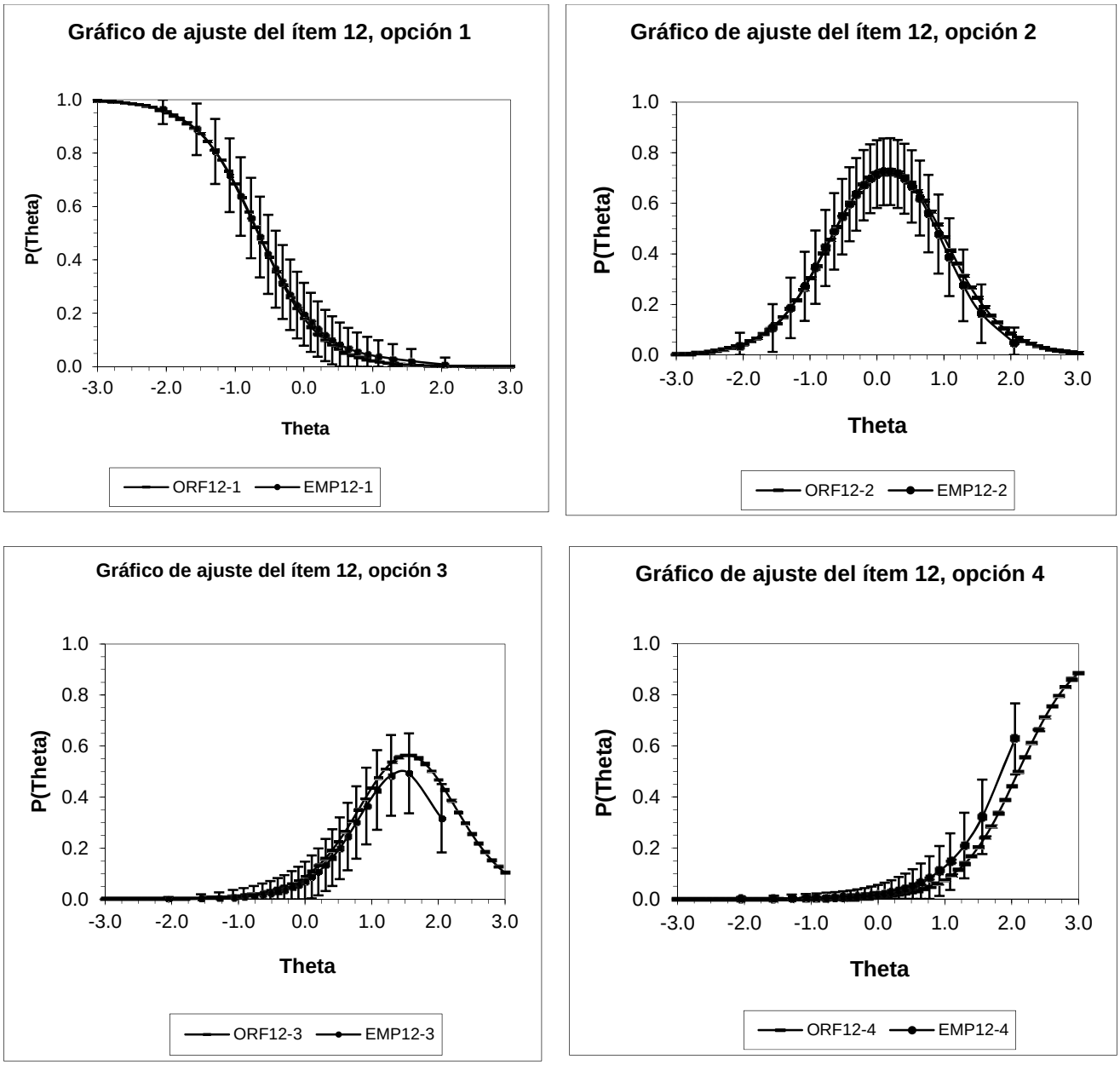



Figura 3.13. Gráfico que examina el ajuste del ítem 12 de la subescala Impulso Motor del BIS al MRG de Samejima, representando para cada opción de respuesta la CCR teórica (ORF) y empírica (EMP).

De los 28 pares de ítems comparados solo 3 exceden el punto de corte fijado (<3), los conjuntos formados por los ítems 2-12, 9-12 y 12-18 con valores de 5.239, 4.492 y 4.770 respectivamente (ver Tabla 3.11). Cabe destacar que estos tres pares de ítems tienen en

común el ya comentado ítem 12. La media de $\chi^2$/g.l. para los conjuntos de dos ítems fue 1.986.

Ninguno de los 56 conjuntos de ítems tomados de tres en tres presenta un ajuste inapropiado, estando todos ellos dentro del rango recomendado (0-3).

Teniendo en cuenta ambos tipos de análisis (estadístico y gráfico), se considera apropiado el ajuste de la subescala de Impulso Motor al MRG de Samejima. La Tabla 3.11 recoge todos los valores de $\chi^2$ calculados.

Tabla 3.11. *Frecuencias de los valores que presenta el estadístico $\chi^2$ dividido por los grados de libertad para cada ítem de la subescala Impulso Motor, así como para conjuntos de dos y tres ítems*

| | $\chi^2$ / g.l | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **<1** | **1<2** | **2<3** | **3<4** | **4<5** | **5<7** | **>7** | **Media** | **DT** |
| **Ítems individuales** | 6 | 0 | 1 | 0 | 0 | 1 | 0 | 1.473 | 1.868 |
| **Conj. 2 ítems** | 3 | 15 | 7 | 0 | 2 | 1 | 0 | 1.986 | 1.154 |
| **Conj. 3 ítems** | 4 | 48 | 4 | 0 | 0 | 0 | 0 | 1.484 | 0.373 |

### 3.1.4.2.2. Subescala Impulso No Planificado del BIS

Se evaluó el ajuste de los parámetros de los ítems calculados en la muestra de calibración con las puntuaciones obtenidas de los sujetos en la muestra de validación, de acuerdo con el MRG de Samejima. Tanto el ajuste gráfico como el ajuste estadístico de los tres tipos de índices de $\chi^2$ fue apropiado.

En el caso del índice de bondad de ajuste $\chi^2$ calculado para cada ítem de la subescala de Impulso No Planificado, tan solo el ítem 17 presenta un valor del estadístico ligeramente superior al valor máximo recomendado ($\chi^2$/gl = 3.203), presentando la mayoría de los ítems valores muy por debajo de los recomendados (ver Tabla 3.12).

Observando el ajuste gráfico del ítem 17 (ver Figura 3.14), se puede apreciar que, en líneas generales el ajuste entre las curvas teóricas y empíricas de las categorías es menor en los niveles intermedios de Impulso No Planificado. La categoría 3 del ítem (bastantes veces), sin embargo, presenta un adecuado ajuste en la zona intermedia de $\theta$, siendo los niveles extremos del nivel de rasgo (especialmente el superior) el que presenta cierto desajuste.

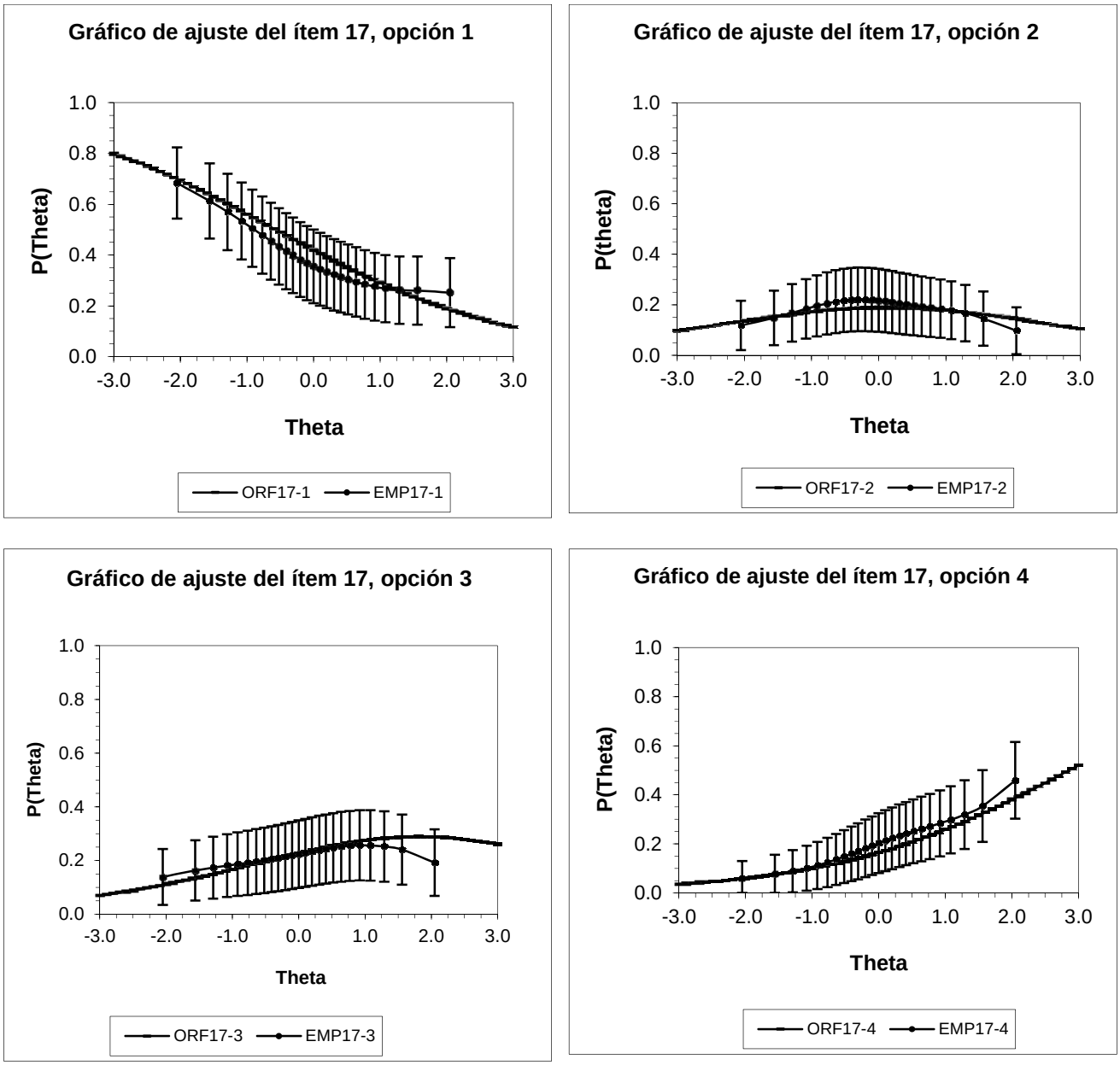



Figura 3.14. Gráfico que examina el ajuste del ítem 17 de la subescala Impulso Motor del BIS al MRG de Samejima, representando para cada opción de respuesta la CCR teórica (ORF) y empírica (EMP).

En el caso de los pares de ítems, también hay únicamente un conjunto de dos ítems (14-17) que presenta unos valores de ajuste ligeramente por encima del criterio especificado ($\chi^2$/gl = 3.310). La media de $\chi^2$/g.l. de los 36 pares de ítems fue 1.653 (ver Tabla 3.12).

En el caso de grupos de tres ítems, la media de $\chi^2$/g.l. fue 1.435, estando los 84 conjuntos de ítems analizados en el rango recomendado de 0 a 3.

En la Tabla 3.12 se pueden visualizar todos los valores medios de $\chi^2$ calculados, así como su frecuencia en cada rango de valores, para cada ítem y para cada conjunto de dos y tres ítems. Los resultados apuntan a que es adecuada la utilización del MRG de Samejima en la subescala de Impulso No Planificado del BIS.

Tabla 3.12. *Frecuencias de los valores que presenta el estadístico $\chi^2$ dividido por los grados de libertad para cada ítem de la subescala Impulso No Planficado, así como para conjuntos de dos y tres ítems*

| | $\chi^2$ / g.l | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **<1** | **1<2** | **2<3** | **3<4** | **4<5** | **5<7** | **>7** | **Media** | **DT** |
| **Ítems individuales** | 5 | 2 | 1 | 1 | 0 | 0 | 0 | 1.189 | 1.003 |
| **Conj. 2 ítems** | 8 | 18 | 9 | 1 | 0 | 0 | 0 | 1.653 | 0.695 |
| **Conj. 3 ítems** | 11 | 65 | 8 | 0 | 0 | 0 | 0 | 1.435 | 0.409 |

### 3.1.4.2.3. Subescala Impulso Cognitivo-Atencional del BIS

El ajuste estadístico de los ítems individuales es muy bueno, encontrándose solo el ítem 7 fuera del rango deseado con un valor de ($\chi^2$/gl = 3.823). La media de todos los ítems fue de 1.671 (ver Tabla 3.13.). El ajuste gráfico de los ítems de la subescala es adecuado, aunque en el caso del ítem 7 sí se aprecia cierto desajuste entre las curvas teóricas y empíricas de sus categorías (ver Figura 3.15). En particular, las categorías 1 y 2 del ítem muestran cierto desajuste en los niveles inferiores del rasgo mientras que en la categoría 3 esto se da en los niveles superiores y el solapamiento de ambas curvas en la categoría 4 se da en prácticamente todos los niveles de adhesión al ítem.

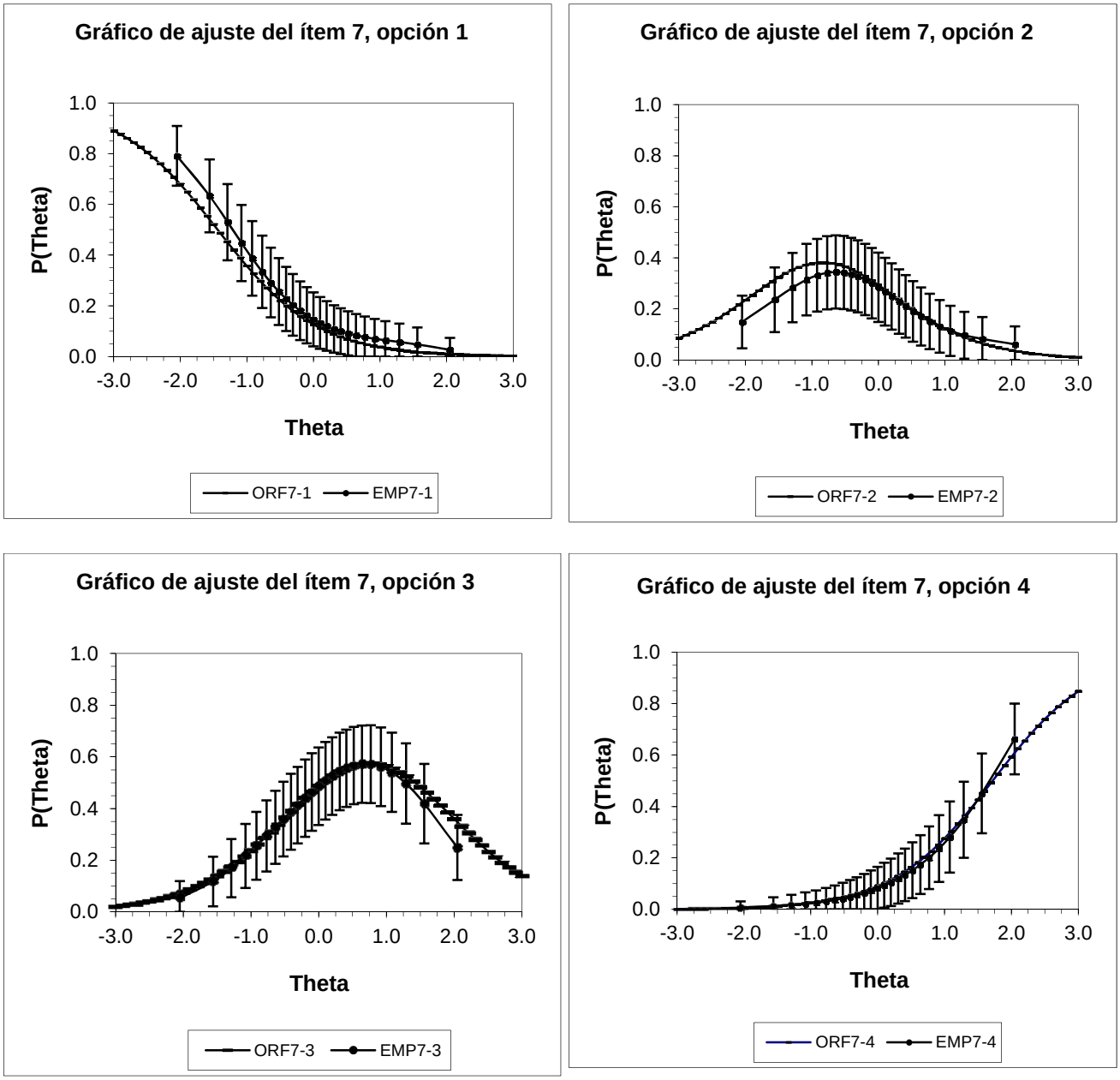



Figura 3.15. Gráfico que examina el ajuste del ítem 7 de la subescala Impulso Cognitivo-Atencional del BIS al MRG de Samejima, representando para cada opción de respuesta la CCR teórica (ORF) y empírica (EMP).

La media de $\chi^2$/gl para los conjuntos de dos ítems fue 2.805, encontrándose 12 de los 45 pares de ítems comparados con un valor de $\chi^2$/gl superior a 3. La mayoría de estos conjuntos de ítems desajustados tienen valores entre tres y cuatro, como es el caso de los

pares 4-7, 4-20, 7-16, 10-13, 10-20, 10-24, 13-16 y 16-19. Sin embargo, hay dos conjuntos de ítems que presentan un valor superior a 7: el formado por los ítems 4-10 ($\chi^2$/gl = 7.765) y el formado por los ítems 10 y 19 ($\chi^2$/gl = 8.996).

Se comparan las 120 combinaciones de tres ítems encontrando que 11 presentan valores desajustados. De éstas había 7 con un valor de $\chi^2$/gl entre 3 y 4 (los conjuntos de ítems 4-7-10, 4-10-13, 4-10-20, 7-10-19, 10-13-19, 10-19-21 y 10-19-27), 3 cuyo con un valor comprendido entre 4 y 5 (los conjuntos 10-16-19 y 10-19-20) y uno con un valor entre 5 y 6 (el conjunto 4-10-19).

En la Tabla 3.13 se muestran las frecuencias de $\chi^2$/gl para ítems individuales y para los conjuntos de dos y tres ítems. El ajuste estadístico de los ítems individuales es bueno, al igual que sucede en el resto de las subescalas, empeorando en las comparaciones de los conjuntos de ítems dobles y triples. De todas formas, ninguna de las medias de $\chi^2$/gl consideradas es superior a 3 por lo que se considera apropiada la utilización del MRG de Samejima.

Tabla 3.13. *Frecuencias de los valores que presenta el estadístico $\chi^2$ dividido por los grados de libertad para cada ítem de la subescala Impulso Cognitivo-Atencional, así como para conjuntos de dos y tres ítems*

| | $\chi^2$ / g.l | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **<1** | **1<2** | **2<3** | **3<4** | **4<5** | **5<7** | **>7** | **Media** | **DT** |
| **Ítems individuales** | 3 | 3 | 3 | 1 | 0 | 0 | 0 | 1.671 | 1.174 |
| **Conj. 2 ítems** | 0 | 13 | 18 | 8 | 4 | 0 | 2 | 2.805 | 1.522 |
| **Conj. 3 ítems** | 0 | 79 | 30 | 7 | 3 | 1 | 0 | 2.025 | 0.728 |

### 3.1.4.2.4. Escala BIS completa

La media del índice $\chi^2$/g.l. para los ítems individuales fue de 1.585, estando 23 de los 27 ítems del test BIS-PA situados en el rango de ajuste adecuado (<3) según recomendaciones de Drasgow *et al*. (1995) (ver Tabla 3.14.).

Todos los ítems tuvieron valores de $\chi^2$/g.l. por debajo de 4, exceptuando el ítem 7 y el ítem 12 (5.24 y 5.33 respectivamente). El ajuste gráfico de ambos ítems puede verse en las Figuras 3.16 y 3.17, donde se representa para cada categoría de respuesta del ítem, el ajuste entre la CCR teórica (ORF) y empírica (EMP). Los ítems 11 y 27 presentan un valor de $\chi^2$/g.l. ligeramente superior al adecuado (3.18 y 3.01 respectivamente).

En general, el ítem 7 presenta un mayor nivel de ajuste en los niveles medios de impulsividad en comparación con ambos extremos del continuo (ver Figura 3.16). Las opciones de respuesta 1 (nunca o casi nunca) y 2 (algunas veces) presentan un mayor desajuste en niveles bajos de impulsividad que en niveles altos y, sobre todo, intermedios. En la opción de respuesta 3 (bastantes veces) este patrón se invierte, ya que ahora el peor ajuste se produce en los niveles altos de rasgo, siendo muy apropiado en los niveles bajo e intermedio. El ajuste entre la curva teórica (ORF) y empírica (EMP) de la opción de respuesta 4 (siempre o casi siempre) es prácticamente perfecto.

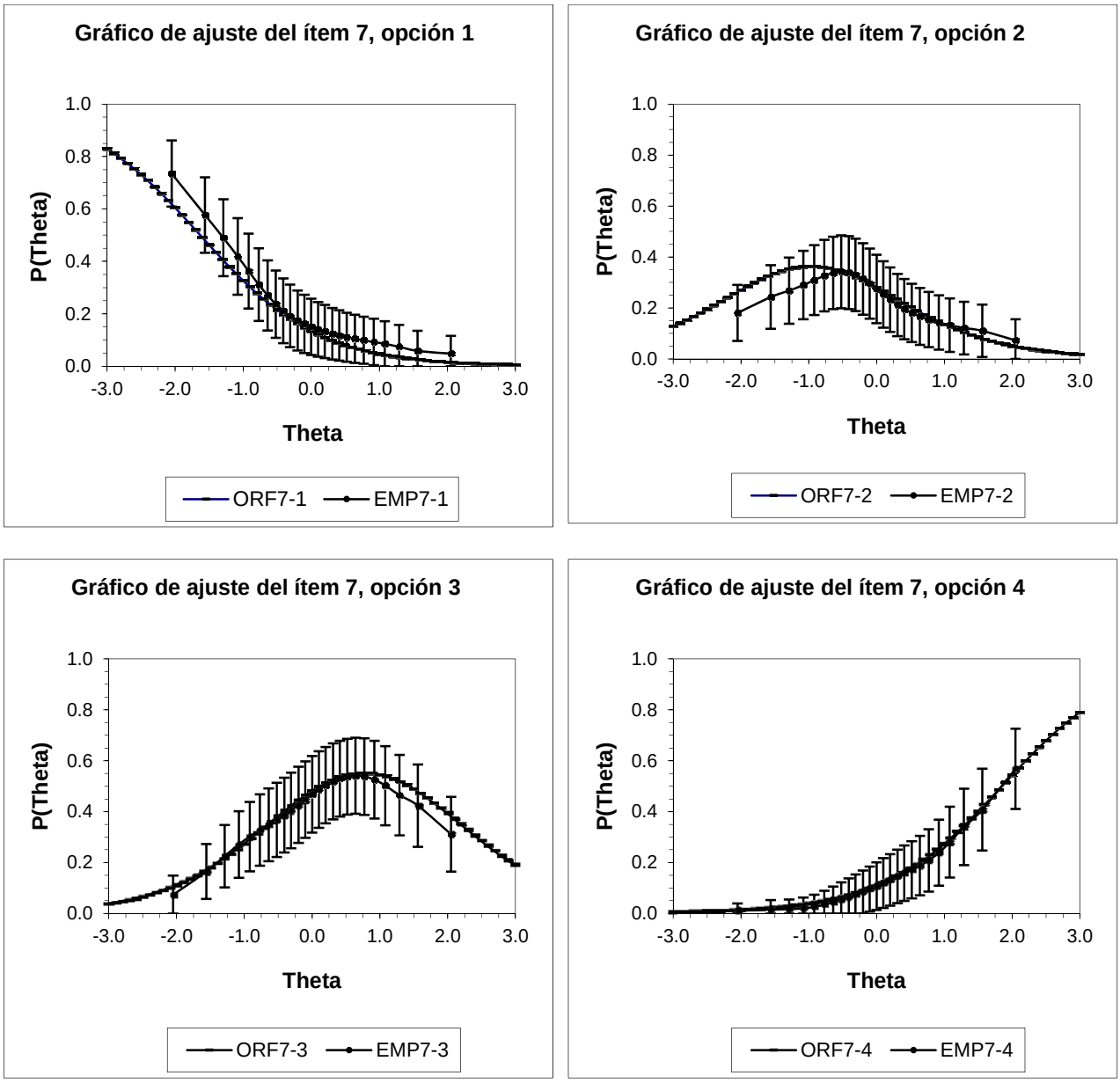



Figura 3.16. Gráfico que examina el ajuste del ítem 7 del BIS al MRG de Samejima, representando para cada opción de respuesta la CCR teórica (ORF) y empírica (EMP).

En el caso del ítem 12, la correspondencia entre la curva teórica (ORF) y empírica (EMP) es peor en niveles altos de impulsividad, habiendo un ajuste apropiado tanto en niveles bajos como intermedios de aptitud, para las cuatro categorías del ítem.

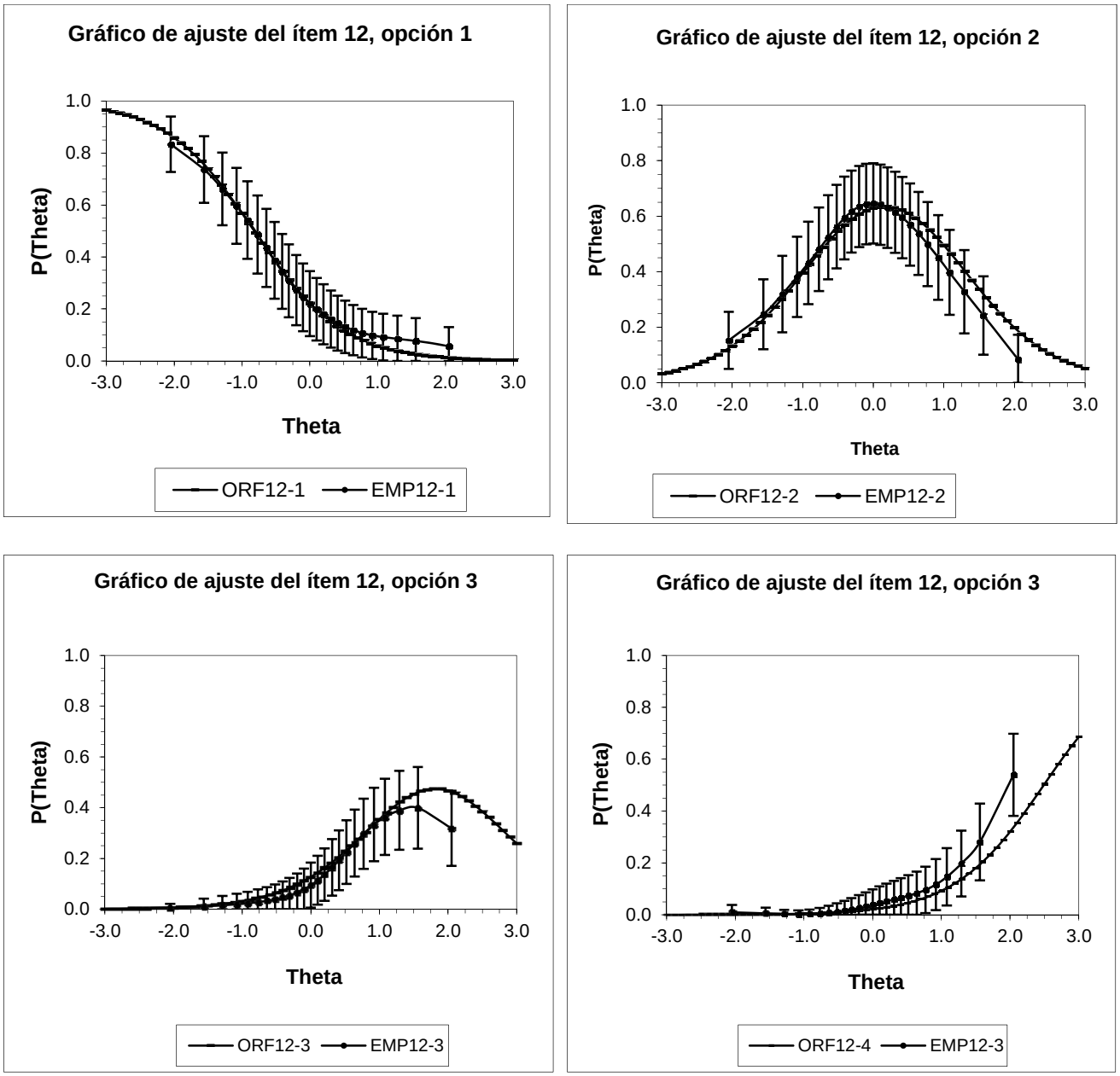



Figura 3.17. Gráfico que examina el ajuste del ítem 12 del BIS al MRG de Samejima, representando para cada opción de respuesta la CCR teórica (ORF) y empírica (EMP).

Se ha representado el ajuste gráfico únicamente de los dos ítems que han obtenido un peor ajuste en términos del estadístico $\chi^2$, por lo que cabe destacar que incluso estos ítems cuyo valor de $\chi^2$ es más desfavorable, presentan un ajuste gráfico que puede considerarse aceptable.

La media de $\chi^2$/g.l. para los conjuntos de dos ítems fue 2.011, con solo tres pares con un valor que excede los criterios recomendados (<3). Este es el caso de los pares de ítems 12-14, 7-16 y 10-18. A excepción de este último par de ítems (10-18) que presentó un valor de $\chi^2$ de 4.19, los otros dos conjuntos de ítems obtuvieron valores de $\chi^2$/g.l. en el rango 3-4 (3.10 y 3.64 respectivamente). En el caso de grupos de tres ítems la media de $\chi^2$/g.l. fue 1.52, estando todos los conjuntos de ítems en el rango de valores recomendados (0-3).

La Tabla 3.14 contiene los valores del estadístico $\chi^2$ para cada ítem individual, así como para los conjuntos de dos y tres ítems. Según estos resultados se puede considerar apropiada la utilización del MRG de Samejima.

Tabla 3.14. *Frecuencias de los valores que presenta el estadístico $\chi^2$ dividido por los grados de libertad para cada ítem del BIS-PA, así como para conjuntos de dos y tres ítems*

| | $\chi^2$ / g.l | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **<1** | **1<2** | **2<3** | **3<4** | **4<5** | **5<7** | **>7** | **Media** | **DT** |
| **Ítems individuales** | 14 | 5 | 4 | 2 | 0 | 2 | 0 | 1.585 | 1.416 |
| **Conj. 2 ítems** | 2 | 12 | 10 | 2 | 1 | 0 | 0 | 2.011 | 0,833 |
| **Conj. 3 ítems** | 1 | 7 | 1 | 0 | 0 | 0 | 0 | 1.515 | 0,423 |

### 3.1.5. RESUMEN DE RESULTADOS

En el marco de la validez se han considerado tres posibles estructuras factoriales para la escala, encontrando que la estructura tridimensional presenta un mejor ajuste a los datos, aunque los datos que consideran la unidimensionalidad y bidimensionalidad de la escala también son aceptables. Se ha analizado la validación cruzada de la muestra

mediante AFC, encontrando un buen ajuste entre muestra de calibración y validación, lo que constituye una evidencia de la generalizabilidad del modelo.

El análisis de ítems se ha llevado a cabo desde el ámbito de la TCT y de la TRI. Según el modelo clásico se ha analizado la discriminación de los ítems mediante correlación ítem-test, presentando valores que oscilan entre .14 y .54 en la discriminación del ítem respecto a su subescala y a la escala total. Estos valores de discriminación moderada son habituales en escalas de personalidad (Ferrando, 1996b). Se han calibrado los ítems mediante el MRG de Samejima, obteniendo unos valores adecuados para los parámetros, que oscilan entre 0.34 y 2.30 en discriminación (parámetro *a*).

La fiabilidad también se ha abordado desde distintas perspectivas: procedimiento clásico (alfa), procedimientos factoriales y procedimiento basado en la TRI. Los análisis clásicos mediante alfa presentan valores entre .63 y .83. Los procedimientos factoriales tienen en cuenta, además, el carácter ordinal de los elementos del test, por lo que obtienen valores mayores de fiabilidad, que oscilan entre .74 y .87 en el caso de alfa ordinal, .77 y .88 según theta y .80 y .93 utilizando la fiabilidad por componentes del AFC (omega). Las funciones de información del MRG de Samejima indican que, en general, los instrumentos arrojan un nivel de información razonable en todos los niveles de impulsividad, siendo más informativos en niveles intermedios y altos de impulsividad.

Por último se evaluó el ajuste en la muestra de validación del MRG de Samejima a los datos, una vez comprobado que los instrumentos poseen la unidimensionalidad necesaria para realizar este tipo de análisis. Esta valoración se realizó teniendo en cuenta

información de tipo estadístico y de tipo gráfico, encontrando en ambos casos un ajuste apropiado del modelo.

Se considera que las aproximaciones utilizadas para valorar la calidad psicométrica tanto de la escala BIS como de sus subescalas arrojan unos resultados más que óptimos, que permiten abordar los estudios de equivalencia que se detallan a continuación.

## 3.2. IMPACTO

Para conocer las diferencias reales en impulsividad medida mediante la escala BIS y sus subescalas, en función de las dos variables estudiadas (sexo y edad) se realiza un MANOVA 2 × 2. Se han asignado pesos diferentes a los sujetos de la muestra, en función de las variables de estratificación, (ver Tabla 2.6) con el fin de otorgar una mayor o menor importancia relativa a algunas unidades muestrales en el análisis estadístico. Esta ponderación está motivada por el hecho de haber utilizado un procedimiento de asignación no proporcional de la muestra a los diferentes estratos, cuyos motivos se han explicado en el apartado 2.1.

### 3.2.1. DIFERENCIAS EN IMPULSIVIDAD EN FUNCIÓN DE LA VARIABLE SEXO

En relación con la variable sexo se encontraron diferencias significativas tanto en el total de la escala $F(1, 1363) = 12.92$ como en las distintas subescalas, con una significación $p < .01$ en el total de la escala y en la subescala Impulso No Planificado $F(1, 1363) = 20.08$, y con una significación $p < .05$ en la subescala de Impulso Motor $F(1, 1363) = 4.22$

y en la subescala Impulso Cognitivo-Atencional $F(1, 1363) = 4.24$, mostrando los chicos una mayor impulsividad en estas medidas que las chicas. Sin embargo, el estadístico *F* está muy influenciado por el tamaño muestral y en los casos como el actual, en el que hay un elevado nº de personas, se obtienen con facilidad resultados estadísicamente significativos (Cohen, 1988). El tamaño del efecto de estas diferencias es muy pequeño, con valores de $\eta^2$ desde .003 (en las subescalas Impulso Motor e Impulso Cognitivo-Atencional) hasta .015 en la subescala de Impulso No Planificado). En la Tabla 3.15 se presentan los estadísticos descriptivos obtenidos por hombres y mujeres, tanto para cada una de las subescalas como para el total del test BIS.

Tabla 3.15. *Medias y desviaciones típicas de la escala global y subescalas del BIS-PA, en el total de la muestra y desglosadas por sexo*

| INSTRUMENTOS: | Hombres | | Mujeres | | Total | |
|---|---|---|---|---|---|---|
| | $\overline{X}$ | $S_x$ | $\overline{X}$ | $S_x$ | $\overline{X}$ | $S_x$ |
| Total BIS-PA (27 ítems) | 2.08 | 0.36 | 2.03 | 0.38 | 2.05 | 0.37 |
| I. Motor (8 ítems) | 1.95 | 0.44 | 1.91 | 0.49 | 1.93 | 0.47 |
| I. No Planificado (9 ítems) | 2.08 | 0.40 | 2.00 | 0.41 | 2.04 | 0.41 |
| I. Cognitivo-Atencional (10 ítems) | 2.24 | 0.50 | 2.20 | 0.49 | 2.22 | 0.49 |

Nota: las puntuaciones oscilan entre 1 y 4. A mayor puntuación mayor nivel del rasgo

### 3.2.2. DIFERENCIAS EN IMPULSIVIDAD EN FUNCIÓN DE LA VARIABLE EDAD

En relación con la variable edad se encontraron diferencias significativas en el total de la escala $F(1, 1363) = 135.69$, $p < .01$, así como en las tres subescalas: Impulso Motor $F(1, 1617) = 69.58$, $p < .01$, Impulso No Planificado $F(1, 1363) = 130.34$, $p < .01$, e Impulso Cognitivo-Atencional $F(1, 1363) = 71.92$, $p < .01$, mostrando en todos los casos unos mayores niveles de impulsividad los adolescentes. Sin embargo el tamaño del efecto también es muy pequeño, tanto en el test completo ($\eta^2 = .091$) como en las tres subescalas

por separado (con valores de $\eta^2$ igual a .049, .087 y .050 para Impulso Motor, Impulso No Planificado e Impulso Cognitivo-Atencional respectivamente).

Tabla 3.16. *Medias y desviaciones típicas de la escala global y subescalas del BIS-PA, en el total de la muestra y desglosadas por edad*

| INSTRUMENTOS: | Preadolescentes | | Adolescentes | | Total | |
|---|---|---|---|---|---|---|
| | $\bar{X}$ | $S_x$ | $\bar{X}$ | $S_x$ | $\bar{X}$ | $S_x$ |
| Total BIS-PA (27 ítems) | 1.93 | 0.34 | 2.16 | 0.37 | 2.05 | 0.38 |
| I. Motor (8 ítems) | 1.81 | 0.44 | 2.03 | 0.47 | 1.93 | 0.47 |
| I. No Planificado (9 ítems) | 1.91 | 0.35 | 2.15 | 0.42 | 2.04 | 0.41 |
| I. Cognitivo-Atencional (10 ítems) | 2.10 | 0.49 | 2.32 | 0.47 | 2.22 | 0.49 |

Nota: las puntuaciones oscilan entre 1 y 4. A mayor puntuación mayor nivel del rasgo

### 3.2.3. DIFERENCIAS EN IMPULSIVIDASD EN FUNCIÓN DE LA INTERACCIÓN EDAD/ SEXO

La interacción de ambas variables (sexo y edad) no arrojó resultados significativos en la escala total BIS, $F(1, 1363) = .84$, con $p = .36$, ni en las subescalas Impulso no planificado e Impulso Cognitivo-Atencional: $F(1, 1363) = .71$, $p = .40$ y $F(1, 1363) = .02$, $p = .89$, respectivamente. Sí hay un leve efecto de la interacción en la subescala Impulso motor $F(1, 1363) = 8.24$, $p < .01$, aunque con un bajo $\eta^2$ (igual a .006), presentando los chicos en la preadolescencia menores niveles de Impulso Motor que las chicas, que tiende a invertirse en la adolescencia (ver Figura 3.18).

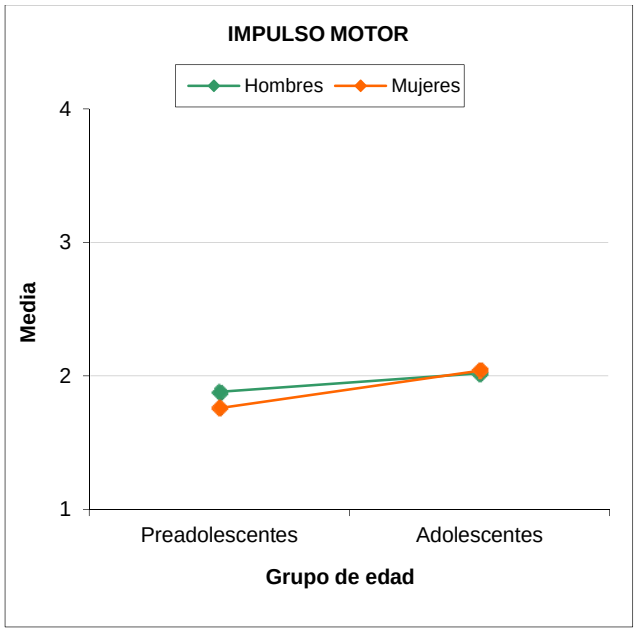



Figura 3.18. Interacción de las variables sexo y edad en la subescala IM.

## 3.3. INVARIANZA MEDIANTE AFC MULTIGRUPO

### 3.3.1. EQUIVALENCIA DE MEDIDA ENTRE HOMBRES Y MUJERES

Los resultados obtenidos mediante la validación cruzada de la muestra fueron satisfactorios, lo que permite utilizar el total de los sujetos (muestra de calibración y muestra de validación) para los estudios de equivalencia. Para probar la equivalencia de medida entre ambos sexos se parte de un primer modelo (modelo base), en el que las cargas factoriales y las varianzas fueron estimadas libremente para hombres y mujeres. En el modelo de invarianza métrica total se fuerza la igualdad de cargas factoriales entre hombres y mujeres, comparando el ajuste entre ambos modelos mediante el incremento en $\chi^2$ y en CFI. En el caso de encontrar diferencias significativas, se busca a los ítems causantes del desajuste, poniendo a prueba la equivalencia métrica parcial de medida. Por

último se pone a prueba el modelo de invarianza escalar total forzando la igualdad de todos los interceptos de los ítems (excepto los liberados en el modelo final de equivalencia métrica parcial de medida) y se comprueba el ajuste con respecto al modelo base de manera análoga al modelo anterior. De haber desajuste se liberarán los ítems correspondientes en el ámbito de la equivalencia escalar parcial de medida.

Para asegurar la identificación del modelo es necesario fijar la carga factorial de un ítem por subescala. En este caso se fijaron a 1 las cargas de los ítems 2, 11 y 16, que pertenecen a las subescalas IM, INP e ICA, respectivamente.

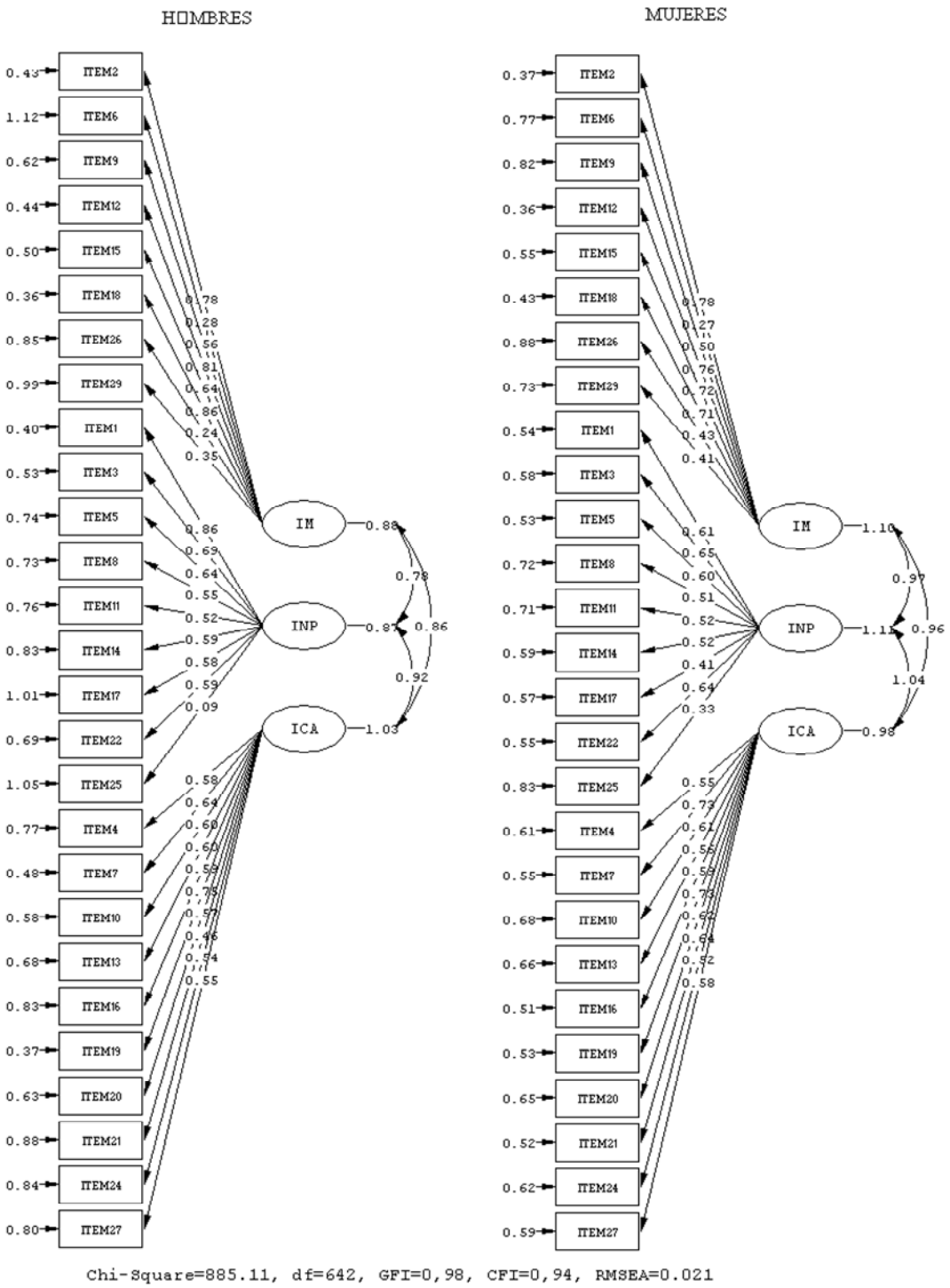



Figura 3.19. Diagrama de vías del AFC multigrupo (hombres y mujeres) del modelo base.

El valor de $\chi^2$ para este modelo base fue 885.11 con 642 grados de libertad. El valor de CFI fue .94 y el de RMSEA fue .021. Estos índices de bondad de ajuste, entre otros,

pueden verse en la Tabla 3.19, e indican que este modelo trifactorial representa de manera adecuada tanto a hombres como a mujeres. Además de comprobar el ajuste global del modelo se examinan los valores estimados para las cargas factoriales en ambos sexos, que tienen valores razonables (ver Tabla 3.17), a excepción del ítem 25, cuyas cargas factoriales son .09 y .33 para hombres y mujeres respectivamente.

Tabla 3.17. *Cargas factoriales estimadas para ambos sexos del modelo base*

| | IM | | INP | | ICA | |
|---|---|---|---|---|---|---|
| | Hombres | Mujeres | Hombres | Mujeres | Hombres | Mujeres |
| Ítem 2 | .78 | .78 | | | | |
| Ítem 6 | .28 | .27 | | | | |
| Ítem 9 | .56 | .50 | | | | |
| Ítem 12 | .81 | .76 | | | | |
| Ítem 15 | .64 | .72 | | | | |
| Ítem 18 | .86 | .71 | | | | |
| Ítem 26 | .24 | .43 | | | | |
| Ítem 29 | .35 | .41 | | | | |
| Ítem 1 | | | .86 | .61 | | |
| Ítem 3 | | | .69 | .65 | | |
| Ítem 5 | | | .64 | .60 | | |
| Ítem 8 | | | .55 | .51 | | |
| Ítem 11 | | | .52 | .52 | | |
| Ítem 14 | | | .59 | .52 | | |
| Ítem 17 | | | .58 | .41 | | |
| Ítem 22 | | | .59 | .64 | | |
| Ítem 25 | | | .09 | .33 | | |
| Ítem 4 | | | | | .58 | .55 |
| Ítem 7 | | | | | .64 | .73 |
| Ítem 10 | | | | | .60 | .61 |
| Ítem 13 | | | | | .60 | .56 |
| Ítem 16 | | | | | .59 | .59 |
| Ítem 19 | | | | | .75 | .73 |
| Ítem 20 | | | | | .57 | .62 |
| Ítem 21 | | | | | .46 | .64 |
| Ítem 24 | | | | | .54 | .52 |
| Ítem 27 | | | | | .55 | .58 |

Nota: Las cargas factoriales se han estandarizado en una métrica común

Las cargas factoriales estimadas en el modelo de invarianza métrica se muestran en la Tabla 3.18.

Tabla 3.18. *Cargas factoriales estimadas del modelo de equivalencia total*

| | IM | INP | ICA |
|---|---|---|---|
| Ítem 2 | .79 | | |
| Ítem 6 | .29 | | |
| Ítem 9 | .52 | | |
| Ítem 12 | .79 | | |
| Ítem 15 | .68 | | |
| Ítem 18 | .77 | | |
| Ítem 26 | .35 | | |
| Ítem 29 | .42 | | |
| Ítem 1 | | .72 | |
| Ítem 3 | | .65 | |
| Ítem 5 | | .61 | |
| Ítem 8 | | .54 | |
| Ítem 11 | | .52 | |
| Ítem 14 | | .53 | |
| Ítem 17 | | .43 | |
| Ítem 22 | | .63 | |
| Ítem 25 | | .23 | |
| Ítem 4 | | | .58 |
| Ítem 7 | | | .69 |
| Ítem 10 | | | .58 |
| Ítem 13 | | | .59 |
| Ítem 16 | | | .58 |
| Ítem 19 | | | .71 |
| Ítem 20 | | | .58 |
| Ítem 21 | | | .58 |
| Ítem 24 | | | .53 |
| Ítem 27 | | | .53 |

Nota: Las cargas factoriales se han estandarizado en una métrica común

El valor de $\chi^2$ para el modelo forzado a mantener la igualdad de cargas factoriales entre hombres y mujeres fue 1000.55. El valor de CFI fue .92 y el de RMSEA .025 (ver Tabla 3.19). Los valores de estos índices indican que el modelo trifactorial ajusta de manera apropiada. Sin embargo, el incremento en $\chi^2$ del modelo base al modelo de invarianza métrica total fue de 115.44 con 24 grados de libertad, cambio que es significativo a nivel estadístico. Asimismo, el incremento en CFI es .02, superior al criterio de .01 establecido por Cheung y Rensvold (2002), por lo que ambos índices apuntan a que no hay equivalencia métrica total entre hombres y mujeres.

Tabla 3.19. *Índices de bondad de ajuste del modelo base y el modelo forzado a mantener la igualdad de las cargas factoriales*

| MODELO 1 (BASE) | MODELO 2 |
|---|---|
| $\chi^2$ (642) =885.11 | $\chi^2$ (666) =1000.55 |
| GFI = .98 | GFI = .98 |
| NNFI = .94 | NNFI = .92 |
| CFI =.94 | CFI =.92 |
| IFI = .95 | IFI = .92 |
| RMSEA =.021 | RMSEA =.025 |

Por tanto, no se puede afirmar que el modelo base ajuste igual de bien que el modelo con restricciones para el global de la escala BIS. Si la hipótesis de equivalencia total se rechaza, como sucede aquí, es necesario realizar los análisis pertinentes para identificar si hay un conjunto de ítems que sean invariantes entre ambos grupos. Ahora es una cuestión, por tanto, de equivalencia parcial de medida.

Para localizar los ítems que provocan la falta de equivalencia se dejan sin restricción de igualdad de cargas factoriales, uno a uno, a los ítems cuyo desajuste sea

mayor según los índices de modificación. Este procedimiento se da por finalizado cuando se encuentra equivalencia entre el modelo base y el modelo de equivalencia parcial.

En el modelo forzado a mantener la igualdad de las cargas factoriales (modelo 2) hay 9 ítems con índices de modificación significativos. Al ser este número menor de la mitad de los ítems, puede darse la equivalencia parcial de medida (Reise, Widaman y Pugh, 1993). El ítem que presenta un mayor índice de modificación es el ítem 1, por lo que se dejará libre de la imposición de cargas factoriales idénticas entre sexos únicamente este ítem, repitiendo el análisis multimuestra.

Tal y como aparece en el modelo 2P1 de la Tabla 3.20, el ajuste del modelo mejora en gran medida con la eliminación de la restricción del ítem 1, mejorando el ajuste de los índices CFI, NNFI, IFI y RMSEA, cuyo valor en el modelo 2 era de .92 para los tres primeros y .025 para el último, siendo ahora de .93 y .024 respectivamente. En cuanto a los índices que evalúan si esta mejora es suficiente, tenemos que el incremento en $\chi^2$ sigue siendo significativo ($p<.01$), mientras que el incremento en CFI es indicativo de equivalencia, ya que tiene un valor de 0,01 lo que se considera aceptable.

Tabla 3.20. *Índices de bondad de ajuste para los modelos de equivalencia métrica entre grupos (hombres y mujeres)*

| | $\chi^2$ | g.l. | $\Delta\chi^2$ | $\Delta$ g.l. | *p* | GFI | NNFI | CFI | $\Delta$ CFI | RMSEA | Ítems libres |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Modelo Base | 885.11 | 642 | | | | .98 | .94 | .94 | | .021 | Todos |
| Modelo 2T | 1000.55 | 666 | 115.44 | 24 | .01 | .98 | .92 | .92 | .02 | .025 | Ninguno |
| Modelo 2P1 | 972.33 | 665 | 87.22 | 23 | .01 | .98 | .93 | .93 | .01 | .024 | 1 |

Nota: Modelo 2T = modelo de invarianza métrica total; Modelo 2P1= modelo de invarianza métrica parcial, con 1 ítem liberado.

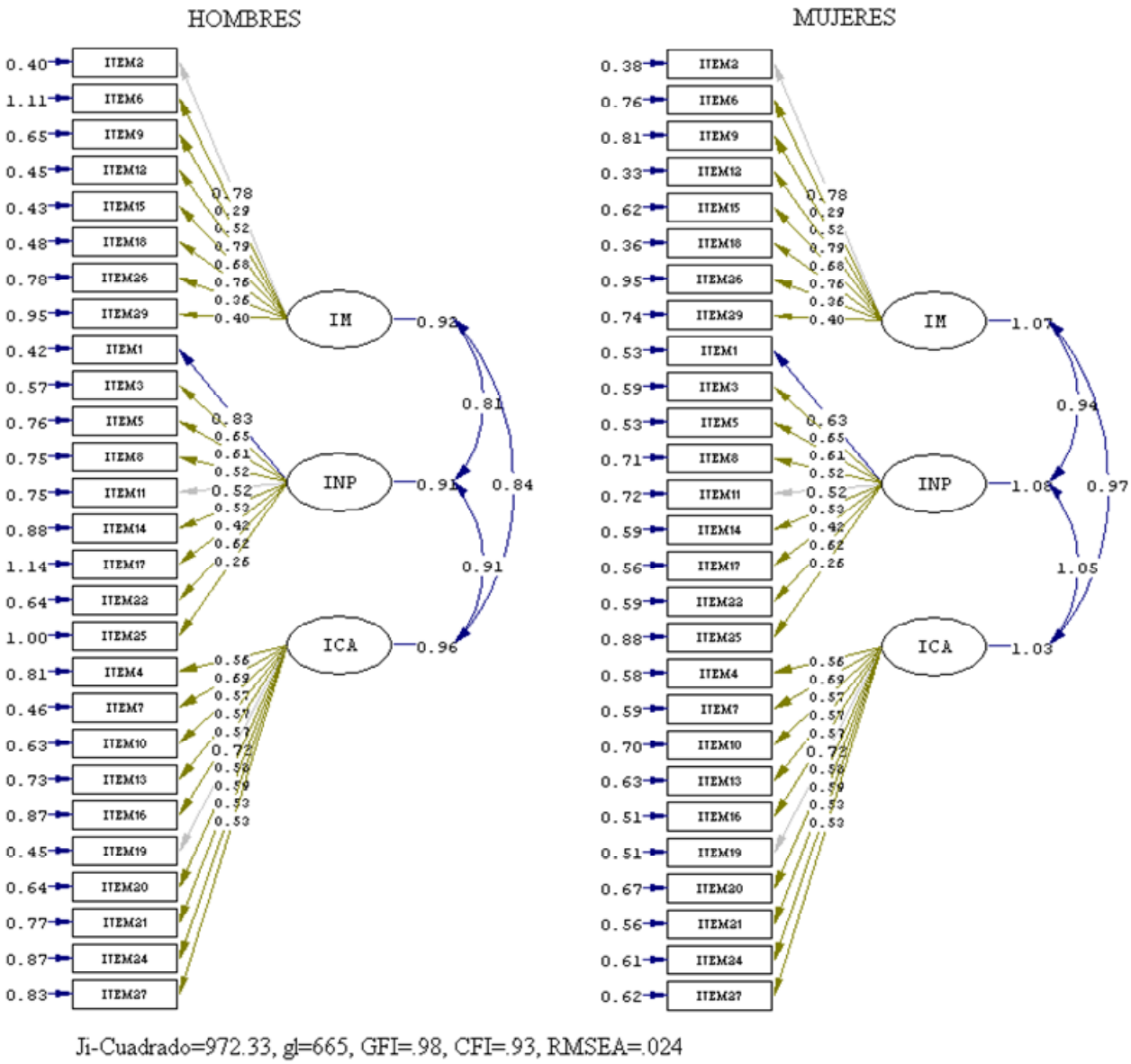
Dado que es conocida la sensibilidad del estadístico $\chi^2$ al tamaño muestral, se considera el criterio del incremento en CFI para considerar este último modelo como definitivo, ya que establece la equivalencia entre hombres y mujeres. Este modelo restringe a la igualdad a todas las cargas factoriales de los ítems a excepción del ítem 1, por lo que hay equivalencia métrica parcial de medida entre ambos grupos.

En la Tabla 3.21 se muestra la estimación de las cargas factoriales para cada ítem forzadas a la igualdad entre sexos, a excepción del ítems que rompe la equivalencia entre grupos. La representación gráfica del modelo de vías multigrupo puede verse en la Figura 3. 20.

Tabla 3.21. *Cargas factoriales estimadas en el modelo 7 de equivalencia parcial*

| | IM | | INP | | ICA | |
|---|---|---|---|---|---|---|
| | Hombres | Mujeres | Hombres | Mujeres | Hombres | Mujeres |
| Ítem 2 | .78 | | | | | |
| Ítem 6 | .29 | | | | | |
| Ítem 9 | .52 | | | | | |
| Ítem 12 | .79 | | | | | |
| Ítem 15 | .68 | | | | | |
| Ítem 18 | .76 | | | | | |
| Ítem 26 | .36 | | | | | |
| Ítem 29 | .40 | | | | | |
| Ítem 1 | | | .82 | .62 | | |
| Ítem 3 | | | .65 | | | |
| Ítem 5 | | | .61 | | | |
| Ítem 8 | | | .52 | | | |
| Ítem 11 | | | .52 | | | |
| Ítem 14 | | | .54 | | | |
| Ítem 17 | | | .45 | | | |
| Ítem 22 | | | .62 | | | |
| Ítem 25 | | | .25 | | | |
| Ítem 4 | | | | | .55 | |
| Ítem 7 | | | | | .68 | |
| Ítem 10 | | | | | .59 | |
| Ítem 13 | | | | | .57 | |
| Ítem 16 | | | | | .57 | |
| Ítem 19 | | | | | .73 | |
| Ítem 20 | | | | | .59 | |
| Ítem 21 | | | | | .59 | |
| Ítem 24 | | | | | .53 | |
| Ítem 27 | | | | | .55 | |

Nota: Las cargas factoriales se han estandarizado en una métrica común




Nota: las flechas grises indican que el parámetro se ha fijado a 1 para asegurar la identificación del modelo, las flechas azules indican que el parámetro varía libremente entre grupos y las flechas verdes indican que se ha forzado la igualdad del parámetro entre grupos.

Figura 3.20. Diagrama de vías del AFC multigrupo (hombres y mujeres) del modelo de invarianza métrica parcial (modelo 2P1).

Una vez comprobada la equivalencia métrica entre ambos sexos se procede a comprobar la equivalencia escalar (igualdad de ordenadas en el origen) con el mismo procedimiento.

Se utiliza para la comparación el mismo modelo base, que deja libertad de parámetros entre los parámetros estimados a hombres y mujeres. El modelo de invarianza escalar completa (modelo 3T) fuerza la igualdad de todos interceptos de los ítems entre los dos grupos estudiados, a excepción del ítem 1 que no mostró garantías de equivalencia métrica. Como se aprecia en la Tabla 3.22, los índices globales de ajuste no son apropiados; asimismo, la comparación de este modelo (modelo 3T) con el modelo base es estadísticamente significativa, $\Delta\chi^2$ =428.02, siendo el incremento en CFI .09, valor muy superior al máximo recomendado, por lo que no puede establecerse la equivalencia escalar total entre ambos grupos.

Para comprobar la equivalencia escalar parcial, se atiende a los índices de modificación para liberar sucesivamente la restricción de igualdad de interceptos de los ítems necesarios. El ítem que presenta un mayor índice de modificación es el ítem 25, por lo que se deja libre de la imposición de igualdad de interceptos de los ítems, repitiendo el análisis multimuestra. Tal y como aparece en el modelo 3P1 (modelo de equivalencia escalar parcial con 1 ítem liberado de la restricción de igualdad) de la Tabla 3.22, hay diferencias importantes entre este modelo y el modelo base en el $\Delta\chi^2$ y el ΔCFI, por lo que se liberan las restricciones del siguiente ítem: el ítem 6. Aunque ahora los índices globales de ajuste se acercan más a valores apropiados, sigue habiendo diferencias significativas según el $\Delta\chi^2$ entre ambos modelos, además de obtenerse un ΔCFI inapropiado, por lo que se continua este proceso. Los siguientes ítems liberados sucesivamente de la restricción de igualdad de interceptos entre grupos son los ítems 8, 29, 13, 24, 27 y 10.

Llegados a este último modelo de equivalencia escalar parcial con 8 ítems liberados de la restricción de igualdad (modelo 3P8), sigue habiendo diferencias significativas en el

$\Delta\chi^2$ pero el valor del incremento en CFI es ya un valor adecuado (0,01), por lo que se finaliza el proceso de modelos anidados.

Tabla 3.22. *Índices de bondad de ajuste para los modelos de equivalencia escalar entre grupos (hombres y mujeres)*

| | $\chi^2$ | g.l. | $\Delta\chi^2$ | $\Delta$ g.l. | *p* | GFI | NNFI | CFI | $\Delta$ CFI | RMSEA | Ítems libres |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Modelo Base | 885.11 | 642 | | | | .98 | .94 | .94 | | .021 | Todos |
| Modelo 3T | 1313.13 | 685 | 428.02 | 43 | .01 | .98 | .85 | .85 | .09 | .033 | Ninguno* |
| Modelo 3P1 | 1212.86 | 684 | 327.75 | 42 | .01 | .98 | .87 | .87 | .07 | .031 | 25 |
| Modelo 3P2 | 1113.04 | 683 | 227.93 | 41 | .01 | .98 | .89 | .90 | .04 | .028 | y 6 |
| Modelo 3P3 | 1069.67 | 682 | 184.56 | 40 | .01 | .98 | .90 | .91 | .03 | .026 | y 8 |
| Modelo 3P4 | 1036.89 | 681 | 151.78 | 39 | .01 | .98 | .91 | .92 | .02 | .025 | y 29 |
| Modelo 3P5 | 1023.37 | 680 | 138.26 | 38 | .01 | .98 | .92 | .92 | .02 | .025 | y 13 |
| Modelo 3P6 | 1015.00 | 679 | 129.89 | 37 | .01 | .98 | .92 | .92 | .02 | .025 | y 24 |
| Modelo 3P7 | 1003.42 | 678 | 118.31 | 36 | .01 | .98 | .92 | .92 | .02 | .024 | y 27 |
| Modelo 3P8 | 990.48 | 677 | 105.37 | 35 | .01 | .98 | .92 | .93 | .01 | .024 | y 10 |

Nota: Modelo 3T = modelo de invarianza escalar total; Modelo 3P1= modelo de invarianza escalar parcial, con 1 ítem liberado, cuando son dos los ítems liberados el modelo es 3P2 y así sucesivamente.
*a excepción del ítem 1. A partir del modelo 3P2, los ítems libres serán el indicado en la celdilla correspondiente más los reflejados en las filas anteriores de la misma columna.

En la variable sexo, por tanto, hay equivalencia escalar parcial de medida. Los interceptos de este modelo de invarianza escalar parcial se muestran en la Tabla 3.23.

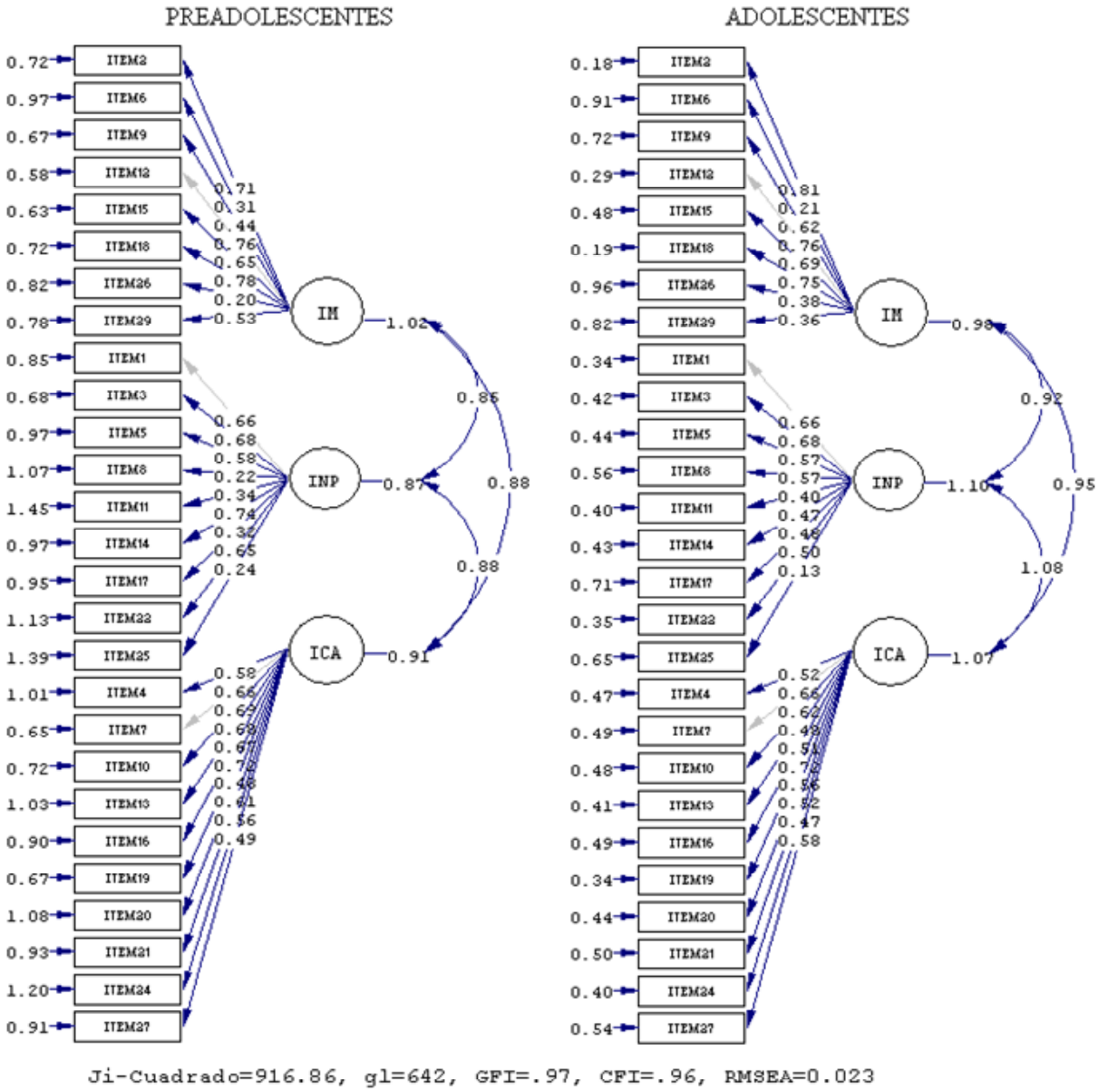
Tabla 3.23. *Interceptos de los ítems estimados para ambos sexos del modelo de invarianza escalar parcial (modelo 3P8)*

| | IM | | INP | | ICA | |
|---|---|---|---|---|---|---|
| | Hombres | Mujeres | Hombres | Mujeres | Hombres | Mujeres |
| Ítem 2 | .00 | .00 | | | | |
| Ítem 6 | -.57 | -.43 | | | | |
| Ítem 9 | -.01 | | | | | |
| Ítem 12 | -.02 | | | | | |
| Ítem 15 | -.16 | | | | | |
| Ítem 18 | -.19 | | | | | |
| Ítem 26 | -.13 | | | | | |
| Ítem 29 | -.24 | .01 | | | | |
| Ítem 1 | | | 1.28 | 1.19 | | |
| Ítem 3 | | | .55 | | | |
| Ítem 5 | | | 1.13 | | | |
| Ítem 8 | | | .63 | .85 | | |
| Ítem 11 | | | .10 | .00 | | |
| Ítem 14 | | | .03 | | | |
| Ítem 17 | | | .42 | | | |
| Ítem 22 | | | .33 | | | |
| Ítem 25 | | | -.49 | -.66 | | |
| Ítem 4 | | | | | .16 | |
| Ítem 7 | | | | | .17 | |
| Ítem 10 | | | | | .15 | -.11 |
| Ítem 13 | | | | | .36 | .48 |
| Ítem 16 | | | | | -.23 | |
| Ítem 19 | | | | | .13 | .00 |
| Ítem 20 | | | | | -.51 | |
| Ítem 21 | | | | | -.50 | |
| Ítem 24 | | | | | -.47 | -.66 |
| Ítem 27 | | | | | -.09 | -.25 |

### 3.3.2. EQUIVALENCIA DE MEDIDA ENTRE PREADOLESCENTES Y ADOLESCENTES

Al igual que en el caso anterior de equivalencia, en el primer modelo (modelo base), las cargas factoriales y las varianzas fueron estimadas libremente para hombres y mujeres.

Las cargas factoriales fijadas a 1 para posibilitar la identificación del modelo son las de los ítems 2, 11 y 16, que pertenecen a las subescalas IM, INP e ICA, respectivamente. En la Figura 3.21 se muestra el diagrama de vías del AFC multigrupo (preadolescentes y adolescentes) del modelo base.




Nota: las flechas grises indican que el parámetro se ha fijado a 1 para asegurar la identificación del modelo y las flechas azules indican que el parámetro varía libremente entre grupos.

Figura 3.21. Diagrama de vías del AFC multigrupo (preadolescentes y adolescentes) del modelo base.

El valor de $\chi^2$ para este modelo "base" fue 916.86 con 642 grados de libertad. Los índices de bondad de ajuste calculados (ver Tabla 3.26) indican que este modelo trifactorial representa de manera adecuada a ambos grupos de edad.

En general, la estimación de los parámetros para ambos grupos de edad tienen valores razonables (ver Tabla 3.24). Atendiendo a la comparativa entre los parámetros estimados a preadolescentes y adolescentes se observa bastante diferencia en algunos casos (ver por ejemplo ítems 8, 14 y 22), lo que podría poner en peligro la equivalencia de cargas factoriales entre ambos grupos.

Tabla 3.24. *Cargas factoriales estimadas para ambos grupos de edad del modelo base*

| | IM | | INP | | ICA | |
|---|---|---|---|---|---|---|
| | Preadolescentes | Adolescentes | Preadolescentes | Adolescentes | Preadolescentes | Adolescentes |
| Ítem 2 | .77 | .77 | | | | |
| Ítem 6 | .33 | .20 | | | | |
| Ítem 9 | .48 | .59 | | | | |
| Ítem 12 | .82 | .73 | | | | |
| Ítem 15 | .70 | .66 | | | | |
| Ítem 18 | .84 | .71 | | | | |
| Ítem 26 | .22 | .36 | | | | |
| Ítem 29 | .58 | .34 | | | | |
| Ítem 1 | | | .73 | .63 | | |
| Ítem 3 | | | .76 | .65 | | |
| Ítem 5 | | | .64 | .54 | | |
| Ítem 8 | | | .24 | .54 | | |
| Ítem 11 | | | .38 | .38 | | |
| Ítem 14 | | | .82 | .44 | | |
| Ítem 17 | | | .36 | .46 | | |
| Ítem 22 | | | .72 | .48 | | |
| Ítem 25 | | | .26 | .13 | | |
| Ítem 4 | | | | | .50 | .59 |
| Ítem 7 | | | | | .57 | .75 |
| Ítem 10 | | | | | .60 | .70 |
| Ítem 13 | | | | | .29 | .54 |
| Ítem 16 | | | | | .58 | .58 |
| Ítem 19 | | | | | .62 | .81 |
| Ítem 20 | | | | | .41 | .64 |
| Ítem 21 | | | | | .53 | .59 |
| Ítem 24 | | | | | .48 | .53 |
| Ítem 27 | | | | | .43 | .65 |

Nota: Las cargas factoriales se han estandarizado en una métrica común

El valor de $\chi^2$ para el modelo forzado a mantener la igualdad de cargas factoriales entre hombres y mujeres (modelo 2) fue 1079.82, con 666 grados de libertad. El valor de CFI fue .94 y el de RMSEA .027 (ver Tabla 3.26). Los valores de estos índices indican que el modelo trifactorial presenta un ajuste apropiado. Sin embargo, el incremento en $\chi^2$ del modelo base al modelo 2 fue de 162.96 con 24 grados de libertad, cambio que es significativo a nivel estadístico. También el decremento en CFI es importante (.2). Las cargas factoriales estimadas en este modelo que fuerza su igualdad se muestran en la Tabla 3.25.

Tabla 3.25. *Cargas factoriales estimadas del modelo de equivalencia total*

| | IM | INP | ICA |
|---|---|---|---|
| Ítem 2 | .79 | | |
| Ítem 6 | .25 | | |
| Ítem 9 | .48 | | |
| Ítem 12 | .75 | | |
| Ítem 15 | .67 | | |
| Ítem 18 | .75 | | |
| Ítem 26 | .29 | | |
| Ítem 29 | .40 | | |
| Ítem 1 | | .65 | |
| Ítem 3 | | .66 | |
| Ítem 5 | | .54 | |
| Ítem 8 | | .45 | |
| Ítem 11 | | .37 | |
| Ítem 14 | | .49 | |
| Ítem 17 | | .38 | |
| Ítem 22 | | .51 | |
| Ítem 25 | | .15 | |
| Ítem 4 | | | .49 |
| Ítem 7 | | | .65 |
| Ítem 10 | | | .62 |
| Ítem 13 | | | .46 |
| Ítem 16 | | | .52 |
| Ítem 19 | | | .71 |
| Ítem 20 | | | .53 |
| Ítem 21 | | | .50 |
| Ítem 24 | | | .48 |
| Ítem 27 | | | .54 |

Nota: Las cargas factoriales se han estandarizado en una métrica común

Tabla 3.26. *Índices de bondad de ajuste del modelo base y el modelo forzado a mantener la igualdad de las cargas factoriales*

| MODELO 1 (BASE) | MODELO 2T |
|---|---|
| $\chi^2$ (642) =916.86 | $\chi^2$ (666) =1079.82 |
| GFI = .97 | GFI = .97 |
| NNFI = .95 | NNFI = .93 |
| CFI =.96 | CFI =.94 |
| IFI = .96 | IFI = .94 |
| RMSEA =.023 | RMSEA =.027 |

En conclusión el ajuste del modelo forzado (modelo 2T) es significativamente peor que el del modelo base, por lo que no hay equivalencia de cargas factoriales entre ambos grupos de edad en el total de la escala. Es el momento de identificar a los ítems que presenten funcionamiento diferencial para comprobar si existe equivalencia parcial y en qué medida.

Para ello, se utiliza un procedimiento no exhaustivo, basado en los índices de modificación para localizar al ítem que presente posible funcionamiento diferencial. Se ejecutará de nuevo el análisis dejando a ese ítem variar libremente entre ambos grupos de edad, volviendo a comparar con el modelo base el nuevo modelo restringido y repitiendo el proceso hasta que no haya diferencias significativas en $\chi^2$ entre ambos modelos.

En el modelo 2T, que restringe la igualdad de las cargas factoriales de todos los ítems de la escala, el ítem que presenta un mayor índice de modificación es el ítem 8, por lo que se repetirá el análisis, sin imponer la igualdad de cargas factoriales idénticas entre ambos grupos de edad únicamente en este ítem (modelo 2P1).

Dado que la mejora del ajuste aún no es suficiente (el incremento en $\chi^2$ sigue siendo significativo y la diferencia en CFI mayor que .01), se eliminará la restricción de igualdad de cargas factoriales del ítem 29, que tiene los valores más altos de modificación en ambos grupos. La repetición del análisis sigue originando incrementos importantes de CFI y de $\chi^2$ por lo que este mismo proceso se repite con el ítem 14 (ver Tabla 3.27).

Tabla 3.27. *Índices de bondad de ajuste para los modelos de equivalencia métrica entre grupos (preadolescentes y adolescentes)*

| | $\chi^2$ | g.l. | $\Delta\chi^2$ | $\Delta$ gl | p | GFI | NNFI | CFI | $\Delta$CFI | RMSEA | Ítems libres |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Modelo Base | 916.86 | 642 | | | | .98 | .94 | .95 | | .023 | Todos |
| Modelo 2T | 1079.82 | 666 | 162.96 | 24 | .01 | .97 | .92 | .92 | .03 | .027 | Ninguno |
| Modelo 2P1 | 1048.40 | 665 | 131.54 | 23 | .01 | .98 | .92 | .93 | .02 | .026 | 8 |
| Modelo 2P2 | 1024.88 | 664 | 108.02 | 22 | .01 | .98 | .93 | .93 | .02 | .026 | y 29 |
| Modelo 2P3 | 1008.14 | 663 | 91.28 | 21 | .01 | .98 | .93 | .94 | .01 | .025 | y 14 |

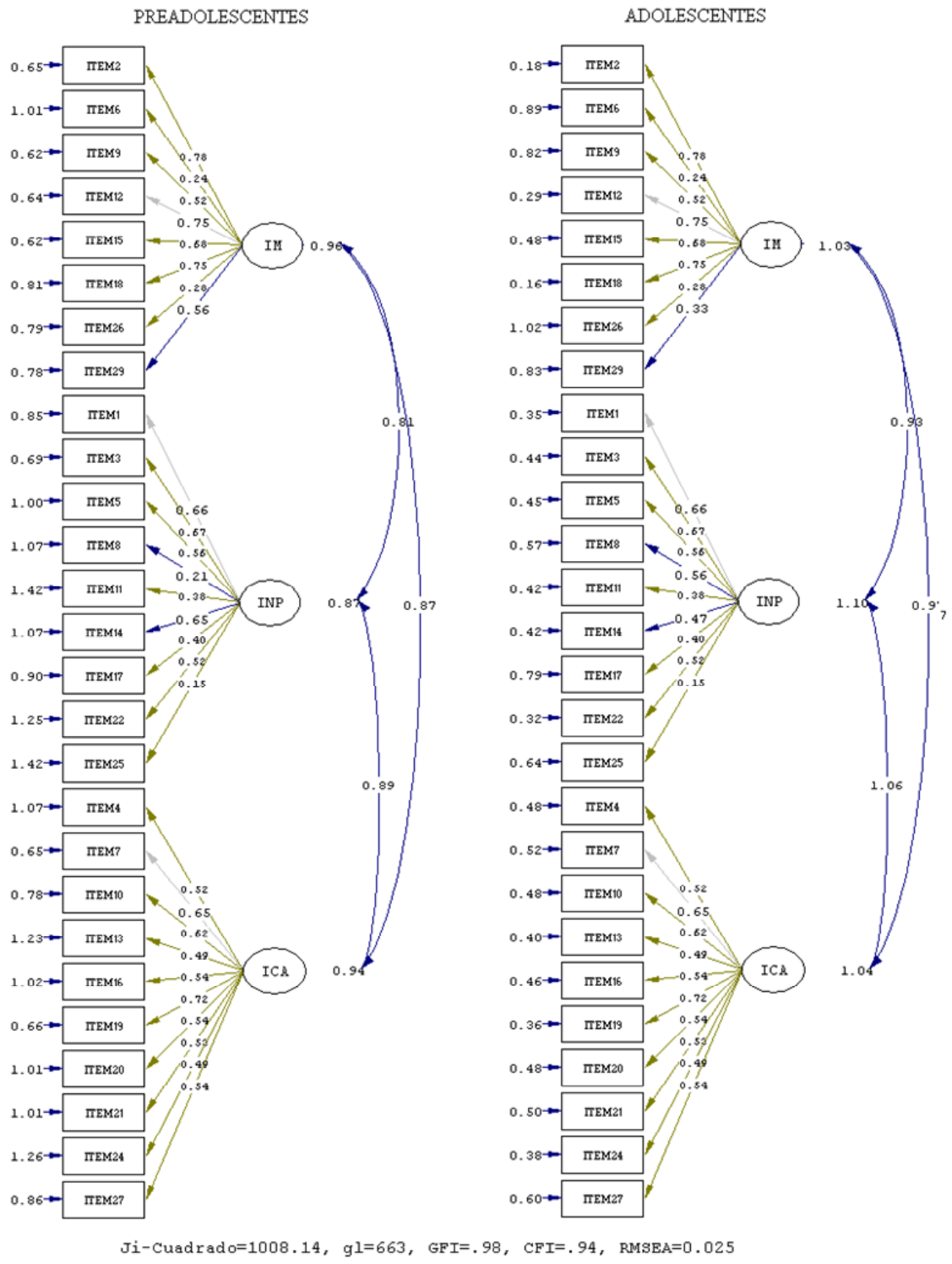
Nota: Modelo 2T = modelo de invarianza métrica total; Modelo 2P1= modelo de invarianza métrica parcial, con 1 ítem liberado, cuando son dos los ítems liberados el modelo es 2P2 y así sucesivamente.
A partir del modelo 2P2, los ítems libres serán el indicado en la celdilla correspondiente (más los reflejados en las filas anteriores de la misma columna).

El último modelo, 2P3 de la Tabla 3.27 (modelo de equivalencia métrica parcial con 3 ítems libres), no arroja diferencias significativas con respecto al modelo base teniendo en cuenta el incremento en CFI, por lo que se considera que existe equivalencia parcial de medida entre preadolescentes y adolescentes. El modelo restringe la igualdad de todas las cargas factoriales de los ítems a excepción de los ítems 8, 29 y 14. La estimación de las cargas factoriales para cada ítem forzadas a la igualdad (a excepción estos tres ítems con DIF) puede verse en la siguiente tabla y en el diagrama de vías de la Figura 3.22.

Tabla 3.28. *Cargas factoriales estimadas en el modelo 2P3 (modelo de equivalencia parcial con 3 ítems libres de restricciones)*

| | IM | | INP | | ICA | |
|---|---|---|---|---|---|---|
| | Preadol. | Adol. | Preadol. | Adol. | Preadol. | Adol. |
| Ítem 2 | .78 | | | | | |
| Ítem 6 | .24 | | | | | |
| Ítem 9 | .52 | | | | | |
| Ítem 12 | .75 | | | | | |
| Ítem 15 | .68 | | | | | |
| Ítem 18 | .75 | | | | | |
| Ítem 26 | .28 | | | | | |
| Ítem 29 | .56 | .33 | | | | |
| Ítem 1 | | | .66 | | | |
| Ítem 3 | | | .67 | | | |
| Ítem 5 | | | .56 | | | |
| Ítem 8 | | | .21 | .56 | | |
| Ítem 11 | | | .38 | | | |
| Ítem 14 | | | .65 | .47 | | |
| Ítem 17 | | | .40 | | | |
| Ítem 22 | | | .52 | | | |
| Ítem 25 | | | .15 | | | |
| Ítem 4 | | | | | .52 | |
| Ítem 7 | | | | | .65 | |
| Ítem 10 | | | | | .62 | |
| Ítem 13 | | | | | .49 | |
| Ítem 16 | | | | | .54 | |
| Ítem 19 | | | | | .72 | |
| Ítem 20 | | | | | .54 | |
| Ítem 21 | | | | | .53 | |
| Ítem 24 | | | | | .49 | |
| Ítem 27 | | | | | .54 | |

Nota: Las cargas factoriales se han estandarizado en una métrica común




Nota: las flechas grises indican que el parámetro se ha fijado a 1 para asegurar la identificación del modelo, las flechas azules indican que el parámetro varía libremente entre grupos y las flechas verdes indican que se ha forzado la igualdad del parámetro entre grupos.

Figura 3.22. Diagrama de vías del AFC multigrupo (preadolescentes y adolescentes) del modelo de invarianza métrica parcial (modelo 2P3).

Para poner a prueba la hipótesis de que además de igualdad de cargas factoriales, existe igualdad de ordenadas en el origen entre los dos grupos, se sigue un procedimiento similar. El mismo modelo base se compara con un modelo de invarianza escalar completa (modelo 3T) que, además de forzar la igualdad de cargas factoriales, obliga a ser iguales a todos los interceptos de los ítems entre preadolescentes y adolescentes, excepto los ítems que presentaron funcionamiento diferencial en el paso anterior (ítems 8, 29 y 14).

Sin necesidad de comparar ambos modelos se observa que el modelo de invarianza escalar completa es claramente inapropiado atendiendo a los índices de ajuste global CFI = .76 y NNFI = .76. Por tanto, se ejecuta un proceso de modelos anidados en el ámbito de la equivalencia parcial, que progresivamente va liberando de restricciones de igualdad de ordenadas en el origen a los ítems cuyos índices de modificación sean mayores.

Este proceso se termina cuando el incremento en $\chi^2$ ya no es significativo o cuando es menor o igual a .01 la diferencia en el índice de ajuste CFI entre el modelo de equivalencia parcial y el modelo base. Como se puede apreciar en la Tabla 3.29 fue necesario liberar de la restricción de igualdad de interceptos a los ítems 10, 11, 25, 5, 13, 3, 21, 24, 6, 19, 27 y 9 (además de los ítems 8, 29 y 14 que presentaron DIF en el paso anterior) para llegar a un modelo que no tuviera diferencias importantes de ajuste con el modelo base.

Tabla 3.29. *Índices de bondad de ajuste para los modelos de equivalencia escalar entre grupos (preadolescentes y adolescentes)*

| | $\chi^2$ | g.l. | $\Delta\chi^2$ | $\Delta$ gl | p | GFI | NNFI | CFI | $\Delta$ CFI | RMSEA | Ítems libres |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Modelo Base | 916.86 | 642 | | | | .98 | .94 | .95 | | .023 | Todos |
| Modelo 3T | 1937.53 | 684 | 1020.1 | 42 | .01 | .94 | .76 | .76 | .19 | .047 | Ninguno* |
| Modelo 3P1 | 1799.33 | 683 | 882.47 | 41 | .01 | .97 | .78 | .79 | .16 | .044 | 10 |
| Modelo 3P2 | 1591.79 | 682 | 674.14 | 40 | .01 | .97 | .82 | .83 | .12 | .040 | y 11 |
| Modelo 3P3 | 1315.48 | 681 | 398.62 | 39 | .01 | .97 | .88 | .88 | .07 | .033 | y 25 |
| Modelo 3P4 | 1240.46 | 680 | 323.60 | 38 | .01 | .97 | .89 | .90 | .05 | .031 | y 5 |
| Modelo 3P5 | 1172.37 | 679 | 255.51 | 37 | .01 | .97 | .90 | .91 | .04 | .030 | y 13 |
| Modelo 3P6 | 1137.13 | 678 | 220.27 | 36 | .01 | .97 | .91 | .91 | .04 | .029 | y 3 |
| Modelo 3P7 | 1111.89 | 677 | 195.03 | 35 | .01 | .97 | .92 | .92 | .03 | .028 | y 21 |
| Modelo 3P8 | 1080.45 | 676 | 163.59 | 34 | .01 | ,97 | .92 | .92 | .03 | .027 | y 24 |
| Modelo 3P9 | 1056.77 | 675 | 139.91 | 33 | .01 | .97 | .93 | .93 | .02 | .026 | y 6 |
| Modelo 3P10 | 1038.89 | 674 | 122.03 | 32 | .01 | .97 | .93 | .93 | .02 | .026 | y 19 |
| Modelo 3P11 | 1027.67 | 673 | 110.81 | 31 | .01 | .97 | .93 | .93 | .02 | .025 | y 27 |
| Modelo 3P12 | 1012.78 | 672 | 95.92 | 30 | .01 | .97 | .93 | .94 | .02 | .025 | y 9 |

Nota: Modelo 3T = modelo de invarianza escalar total; Modelo 3P1= modelo de invarianza escalar parcial, con 1 ítem liberado, cuando son dos los ítems liberados el modelo es 3P2 y así sucesivamente.
*a excepción de los ítems 8, 29 y 14. A partir del modelo 3P2, los ítems libres serán el indicado en la celdilla correspondiente más los reflejados en las filas anteriores de la misma columna.

En el último modelo de invarianza escalar parcial propuesto los únicos ítems cuyas ordenadas en el origen se han forzado a ser iguales son los ítems 2, 12, 15, 18, 26, 1, 17, 22, 4, 7, 16 y 20, mientras que son 15 los ítems sin esta restricción. Por tanto hay más ítems liberados de condición de igualdad de interceptos que ítems que la cumplen, lo que pone en tela de juicio que exista equivalencia escalar parcial en la variable edad. Las ordenadas en el origen de este último modelo de invarianza escalar parcial se muestran en la Tabla 3.30.

Tabla 3.30. *Interceptos de los ítems estimados para preadolescentes y adolescentes del modelo de invarianza escalar parcial (modelo 3P12)*

| | IM | | INP | | ICA | |
|---|---|---|---|---|---|---|
| | Preadol. | Adolescentes | Preadol. | Adolescentes | Preadol. | Adolescentes |
| Ítem 2 | | -.05 | | | | |
| Ítem 6 | -.40 | -.51 | | | | |
| Ítem 9 | .08 | -.05 | | | | |
| Ítem 12 | | .00 | | | | |
| Ítem 15 | | -.16 | | | | |
| Ítem 18 | | -.21 | | | | |
| Ítem 26 | | -.08 | | | | |
| Ítem 29 | -.03 | -.12 | | | | |
| Ítem 1 | | | | .00 | | |
| Ítem 3 | | | -.40 | -.96 | | |
| Ítem 5 | | | .31 | -.16 | | |
| Ítem 8 | | | .13 | -.36 | | |
| Ítem 11 | | | -1.14 | -.63 | | |
| Ítem 14 | | | -1.18 | -.91 | | |
| Ítem 17 | | | | -.41 | | |
| Ítem 22 | | | | -.73 | | |
| Ítem 25 | | | -1.13 | -.73 | | |
| Ítem 4 | | | | | | -.02 |
| Ítem 7 | | | | | | .00 |
| Ítem 10 | | | | | .07 | -.50 |
| Ítem 13 | | | | | .49 | .14 |
| Ítem 16 | | | | | | -.37 |
| Ítem 19 | | | | | -.01 | -.36 |
| Ítem 20 | | | | | | -.71 |
| Ítem 21 | | | | | -.76 | -.59 |
| Ítem 24 | | | | | -.84 | -.70 |
| Ítem 27 | | | | | -.26 | -.46 |

## 3.4. INVARIANZA MEDIANTE COMPARACIÓN DE MODELOS CON EL TEST DE RAZÓN DE VEROSIMILITUD (LR)

Se utiliza el procedimiento de comparación de modelos para evaluar el funcionamiento diferencial del test completo y de cada una de las subescalas, en las dos variables de interés: sexo y edad. Para ello, se compara la verosimilitud de un modelo con restricciones que establece la igualdad del parámetro $a$ o de todos los parámetros de los ítems ($a$, $b_1$, $b_2$ y $b_3$), con la verosimilitud de un modelo base en el que se asume que los parámetros de los ítems del test pueden diferir entre los grupos.

La diferencia entre los valores de verosimilitud de los modelos restringido y base se expresa por medio del estadístico $G^2$ (ver apartado 2.5.4.). Bajo la hipótesis nula, este estadístico sigue una distribución $\chi^2$ con un número de grados de libertad igual a la diferencia entre el número de parámetros estimados en ambos modelos. Si, para un determinado valor de confianza, el valor obtenido es menor que el valor teórico de la distribución, se acepta la hipótesis de no diferencias en el ajuste de ambos, lo que apoyaría la equivalencia de medida de la prueba en la variable estudiada. Si, por el contrario, el valor de $G^2$ es mayor que el valor teórico de la distribución, se rechaza la hipótesis de igualdad en el ajuste de ambos modelos y no se puede hablar de equivalencia total; será necesario identificar a los ítems causantes del desajuste en el marco de la equivalencia parcial de medida.

Para evaluar el DIF se utiliza en este caso el mismo procedimiento de comparación de modelos, con un modelo compacto que establece la igualdad de parámetros en todos los

ítems excepto en el ítem objeto de estudio. Para su estimación se ha utilizado el programa IRTLRDIF.

### 3.4.1. EQUIVALENCIA DE MEDIDA ENTRE HOMBRES Y MUJERES

#### *3.4.1.1. Subescala Impulso Motor del BIS*

Se pone a prueba la igualdad del parámetro de discriminación entre hombres y mujeres, encontrando que no hay diferencias significativas entre el modelo base y el modelo en el que se ha forzado la igualdad de *a* entre hombres y mujeres en todos los ítems de la subescala ($\Delta G^2$ [8] = 13.7, ns). Por tanto, hay equivalencia total de medida entre hombres y mujeres en relación al parámetro de discriminación de los ítems (ver Tabla 3.31).

Tabla 3.31. *Equivalencia de medida entre hombres y mujeres en la subescala de Impulso Motor en relación al parámetro de discriminación*

| | $G^2$ | $\Delta G^2$ | $\Delta$ g.l. | $\chi^2$ | *p* | Ítems libres |
|---|---|---|---|---|---|---|
| Modelo base | 6795.1 | | | | | Todos |
| Modelo invarianza total *a* | 6808.8 | 13.7 | 8 | 20.09 | n.s | Ninguno |

Una vez comprobada este tipo de equivalencia, se pone a prueba un modelo más restrictivo que exige no solo la igualdad de los ítems en el parámetro *a*, sino, además, que no haya variaciones en los parámetros *b* entre ambos grupos.

En este caso sí hay diferencias significativas, ya que $\Delta G^2$ [32] = 69.3, p<.01; por tanto no hay equivalencia completa de medida en ambos parámetros entre sexos en la subescala. Para averiguar qué ítems provocan la falta de equivalencia se analiza el funcionamiento diferencial de los ítems con IRTLRDIF.

En la Tabla 3.32 se presentan los resultados del análisis DIF de los ítems de la subescala Impulso Motor entre hombres y mujeres. En el caso de los ítems que no presentan DIF se presenta únicamente una línea que corresponde con los resultados de poner a prueba la igualdad de los dos grupos en el ítem en todos los parámetros. En el caso de los ítems que sí presentan DIF, además de esta información, se incluye en dos líneas más, los resultados del DIF en el parámetro *a* y en los parámetros *b*. En ambos casos se incluye la estimación de los parámetros para ambos grupos utilizando el MRG de Samejima.

Se han puesto a prueba las hipótesis de igualdad entre ambos grupos en los dos parámetros, en el parámetro *a* y en los parámetros *b*, comparando el valor del estadístico $G^2$ con el valor crítico de la distribución $\chi^2$ utilizando un $\alpha$ =.01, que es 13’28 para 4 grados de libertad, 6.63 para 1 g.l y 11.34 para 3 grados de libertad respectivamente.

Tabla 3.32. *Análisis del funcionamiento diferencial de los ítems de la subescala Impulso Motor entre hombres y mujeres*

| | | | | PARÁMETROS HOMBRES | | | | PARÁMETROS MUJERES | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | Hip | $G^2$ | gl | *a* | $b_1$ | $b_2$ | $b_3$ | *a* | $b_1$ | $b_2$ | $b_3$ |
| 2 | *ab* igual | 1.4 | 4 | 1.97 | -0.71 | 1.24 | 2.52 | 2.05 | -0.65 | 1.21 | 2.61 |
| 6 | *ab* igual | 5.7 | 4 | 0.38 | 1.45 | 3.77 | 5.07 | 0.41 | 0.99 | 3.70 | 4.96 |
| 9 | *ab* igual | 3.1 | 4 | 0.83 | -0.87 | 0.92 | 2.25 | 0.80 | -0.72 | 1.06 | 2.24 |
| 12 | *ab* igual | 4.8 | 4 | 2.22 | -0.60 | 0.96 | 1.96 | 2.63 | -0.62 | 0.92 | 1.96 |
| 15 | *ab* igual | 5.0 | 4 | 1.32 | -0.52 | 1.16 | 2.36 | 1.42 | -0.40 | 0.98 | 2.11 |
| 18 | *ab* igual | 10.9 | 4 | 2.33 | -0.37 | 1.07 | 2.06 | 2.35 | -0.40 | 1.31 | 2.16 |
| 26 | *ab* igual | 11.2 | 4 | 0.39 | -0.43 | 3.85 | 6.87 | 0.55 | -0.51 | 2.65 | 4.11 |
| 29 | *ab* igual | 23.7* | 4 | 0.49 | -0.19 | 2.88 | 5.40 | 0.68 | -0.82 | 1.92 | 3.60 |
| 29 | *a* igual | 2.5 | 1 | 0.61 | -0.14 | 2.39 | 4.46 | 0.61 | -0.91 | 2.14 | 4.02 |
| 29 | *b* igual | 21.2* | 3 | 0.59 | -0.58 | 2.32 | 4.34 | 0.59 | -0.58 | 2.32 | 4.34 |

Nota: *$p < 0.01$

El único ítem que presenta funcionamiento diferencial es el ítem 29, por lo que se elimina la restricción de igualdad de parámetros únicamente en este ítem y se compara nuevamente el modelo base y el modelo con restricciones de igualdad de parámetros. En esta ocasión no hay diferencias significativas entre ambos modelos, dado que $\Delta G^2$ [28] = 45.3, n.s. En la Tabla 3.33. se muestran los resultados derivados de poner a prueba el modelo de invarianza completa de los parámetros *a* y *b*.

Tabla 3.33. *Equivalencia de medida entre hombres y mujeres en la subescala de Impulso Motor forzando la igualdad de todos los parámetros*

| | $G^2$ | $\Delta G^2$ | $\Delta$ g.l. | $\chi^2$ | *p* | Ítems libres |
|---|---|---|---|---|---|---|
| Modelo base | 6795.1 | | | | | Todos |
| Modelo invarianza total *a* y *b* | 6864.4 | 69.3 | 32 | 50.89 | .01 | Ninguno |
| Modelo invarianza parcial *a* y *b* | 6840.4 | 45.3 | 28 | 48.28 | ns | Ítem 29 |

Hay equivalencia parcial de la subescala Impulso Motor entre hombres y mujeres. El único ítem que rompe la equivalencia total entre ambos grupos es el ítem 29. En la Figura 3.23 se muestra la Curva Característica del Test (CCT), en la que se aprecia un gran solapamiento en las curvas de ambos sexos en niveles bajos e intermedios del rasgo, con una puntuación esperada mayor para las mujeres en los niveles altos de impulsividad motora.

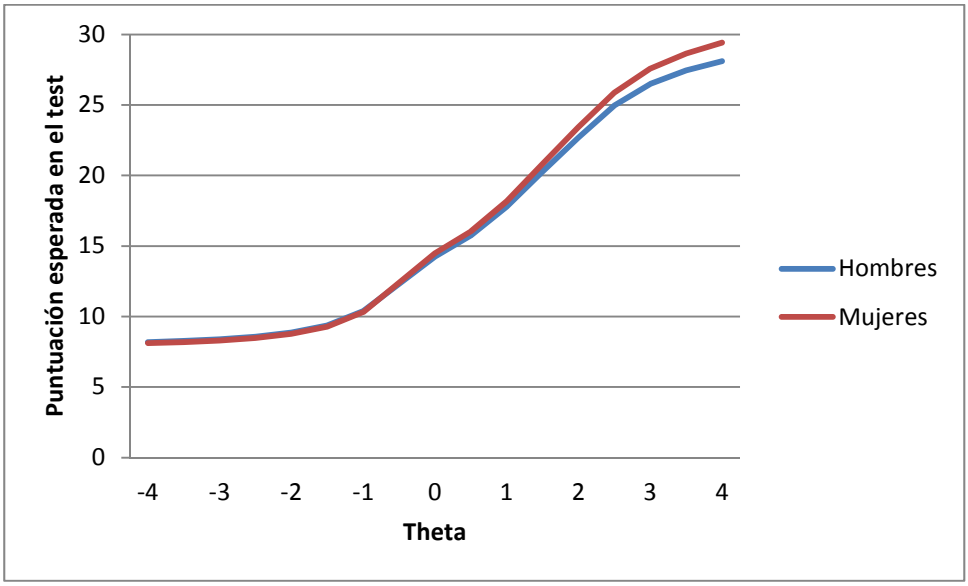



Figura 3.23. CCT para hombres y mujeres en la subescala Impulso Motor.

### *3.4.1.2. Subescala Impulso no Planificado del BIS*

Los resultados indican que no hay invarianza de medida respecto al parámetro *a* de todos los ítems entre el grupo de hombres y de mujeres, ya que el incremento de $G^2$ es significativo ($\Delta G^2$ [9] = 23.7, p<.01).

Mediante el programa IRTLRDIF se analiza el funcionamiento diferencial de todos los ítems de la subescala para localizar a los que rompan la invarianza en el ámbito de la equivalencia parcial de medida entre ambos sexos. En la Tabla 3.34 se indica el DIF relativo a ambos parámetros, desglosándose en cada uno de ellos en los casos en que es significativo, junto con el valor del incremento en $G^2$, los grados de libertad correspondientes, y la estimación realizada para todos los parámetros en ambos grupos.

Tabla 3.34. *Análisis del funcionamiento diferencial de los ítems de la subescala Impulso No Planificado entre hombres y mujeres*

| | | | | PARÁMETROS HOMBRES | | | | PARÁMETROS MUJERES | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | Hip | $G^2$ | gl | *a* | $b_1$ | $b_2$ | $b_3$ | *a* | $b_1$ | $b_2$ | $b_3$ |
| 1 | *ab* igual | 5.1 | 4 | 1.09 | -1.42 | 0.22 | 3.58 | 1.10 | -1.30 | 0.20 | 4.11 |
| 3 | *ab* igual | 11.5 | 4 | 0.94 | -0.96 | 1.73 | 3.62 | 1.19 | -0.54 | 1.77 | 3.18 |
| 5 | *ab* igual | 6.0 | 4 | 0.54 | -2.88 | -0.20 | 2.95 | 0.39 | -4.00 | -0.37 | 4.56 |
| 8 | *ab* igual | 19.3* | 4 | 0.71 | -1.12 | 0.72 | 2.50 | 0.61 | -1.81 | 0.04 | 2.34 |
| 8 | *a* igual | 0.8 | 1 | 0.65 | -1.25 | 0.75 | 2.69 | 0.65 | -1.71 | 0.04 | 2.21 |
| 8 | *b* igual | 18.5* | 3 | 0.60 | -1.61 | 0.38 | 2.57 | 0.60 | -1.61 | 0.38 | 2.57 |
| 11 | *ab* igual | 3.6 | 4 | 1.58 | 0.25 | 1.33 | 2.67 | 1.51 | 0.38 | 1.40 | 2.93 |
| 14 | *ab* igual | 10.7 | 4 | 1.47 | 0.34 | 1.41 | 2.31 | 1.29 | 0.20 | 1.71 | 2.70 |
| 17 | *ab* igual | 23.9* | 4 | 0.70 | -0.47 | 0.52 | 1.99 | 0.44 | -0.89 | 1.01 | 4.11 |
| 17 | *a* igual | 4.2 | 1 | 0.54 | -0.66 | 0.58 | 2.43 | 0.54 | -0.73 | 0.83 | 3.37 |
| 17 | *b* igual | 19.7* | 3 | 0.56 | -0.68 | 0.69 | 2.81 | 0.56 | -0.68 | 0.69 | 2.81 |
| 22 | *ab* igual | 28.1* | 4 | 1.51 | 0.07 | 1.37 | 2.96 | 1.94 | -0.05 | 0.82 | 2.40 |
| 22 | *a* igual | 3.6 | 1 | 1.74 | 0.10 | 1.28 | 2.71 | 1.74 | -0.06 | 0.86 | 2.55 |
| 22 | *b* igual | 24.5* | 3 | 1.68 | -0.01 | 1.05 | 2.64 | 1.68 | -0.01 | 1.05 | 2.64 |
| 25 | *ab* igual | 15.4* | 4 | 0.37 | 2.45 | 5.58 | 8.18 | 0.75 | 1.77 | 3.70 | 5.09 |
| 25 | *a* igual | 6.2 | 1 | 0.57 | 1.76 | 3.86 | 5.59 | 0.57 | 2.24 | 4.73 | 6.54 |
| 25 | *b* igual | 9.2 | 3 | 0.59 | 1.95 | 4.13 | 5.83 | 0.59 | 1.95 | 4.13 | 5.83 |

Nota: *$p < 0.01$

En principio, podrían presentar DIF los ítems 8, 17, 22 y 25. De ellos, el ítem con un valor mayor de $G^2$ para la hipótesis de igualdad de parámetros *a* es el ítem 25, por lo que se anula esta restricción para el ítem antes de comparar de nuevo los modelos.

En esta ocasión la comparación del modelo con restricciones de igualdad en la discriminación de los ítems respecto al modelo base no arroja diferencias significativas ($\Delta G^2$ [8] = 15.5, ns), por lo que hay equivalencia parcial de medida entre hombres y

mujeres. El único ítem que rompe la equivalencia total es el ítem 25. En la Tabla 3.35 se resumen los datos relativos a las pruebas de equivalencia de medida.

Tabla 3.35. *Equivalencia de medida entre hombres y mujeres en la subescala de Impulso No Planificado en relación al parámetro de discriminación*

| | $G^2$ | $\Delta G^2$ | $\Delta$ g.l. | $\chi^2$ | $p$ | Ítems libres |
|---|---|---|---|---|---|---|
| Modelo base | 11039.6 | | | | | Todos |
| Modelo invarianza total *a* | 11063.3 | 21.67 | 9 | 21.67 | .01 | Ninguno |
| Modelo invarianza parcial *a* | 11055.1 | 15.5 | 8 | 20.09 | ns | Ítem 25 |

A continuación se pone a prueba el modelo más restrictivo de igualdad de parámetros, forzando la igualdad de $a$, $b_1$, $b_2$ y $b_3$ entre ambos grupos. Visto que el incremento de $G^2$ es significativo ($\Delta G^2$ [32] = 102, p<.01), se concluye que no hay equivalencia total de medida en la subescala poniendo a prueba la igualdad de todos los parámetros de los ítems (a excepción del ítem 25 cuya igualdad se descartó en el paso anterior) entre hombres y mujeres.

Para comprobar si existe equivalencia parcial se eliminan las restricciones de igualdad de parámetros del ítem 22, por ser el que presenta mayor DIF (ver Tabla 3.34), para comparar de nuevo ambos modelos. Ahora $\Delta G^2$ [28] = 74, p<.01, por lo que sigue sin haber equivalencia entre ambos modelos. El siguiente ítem del que se eliminan sus restricciones de igualdad de parámetros es el ítem 17, tras lo cual, $\Delta G^2$ [24] = 52.9, p<.01. Al no existir igualdad entre los modelos comparados se elimina la restricción de igualdad del ítem 8, para comparar de nuevo los modelos.

En esta ocasión no hay diferencias entre el modelo base y el modelo con restricciones de igualdad de parámetros $\Delta G^2$ [20] = 32.1, n.s., por lo que se concluye equivalencia parcial de medida entre hombres y mujeres.

Es necesario eliminar la restricción de igualdad de parámetros de los cuatro ítems que presentan DIF en la subescala Impulso No Planificado para encontrar equivalencia de medida entre ambos grupos. Estos ítems son: el ítem 25, el ítem 22, el ítem 17 y el ítem 8. En la siguiente Tabla se resumen estos resultados.

Tabla 3.36. *Equivalencia de medida entre hombres y mujeres en la subescala de Impulso No Planificado forzando la igualdad de todos los parámetros*

| | $G^2$ | $\Delta G^2$ | $\Delta$ g.l. | $\chi^2$ | $p$ | Ítems libres |
|---|---|---|---|---|---|---|
| Modelo base | 11039.6 | | | | | Todos |
| Modelo invarianza total *a* y *b* | 11141.6 | 102 | 32 | 50.89 | .01 | Ítem 25 |
| Modelo invarianza parcial *a* y *b* | 11113.6 | 74 | 28 | 48.28 | .01 | Ítems 25 y 22 |
| Modelo invarianza parcial *a* y *b* | 11092.5 | 52.9 | 24 | 42.98 | .01 | Ítems 25, 22 y 17 |
| Modelo invarianza parcial *a* y *b* | 11071.7 | 32.1 | 20 | 37.57 | .01 | Ítems 25, 22, 17 y 8 |

Las CCT para ambos sexos pueden verse en la Figura 3.24. Resulta llamativo que, a pesar de que casi la mitad de los ítems de la subescala (4 ítems) presenta DIF, apenas se aprecian diferencias entre chicos y chicas en la puntuación esperada a nivel de la subescala. Como se tendrá ocasión de comprobar al examinar los resultados con el modelo DFIT (ver apdo. 3.5.), el gráfico puede estar revelando un efecto de compensación entre ítems.

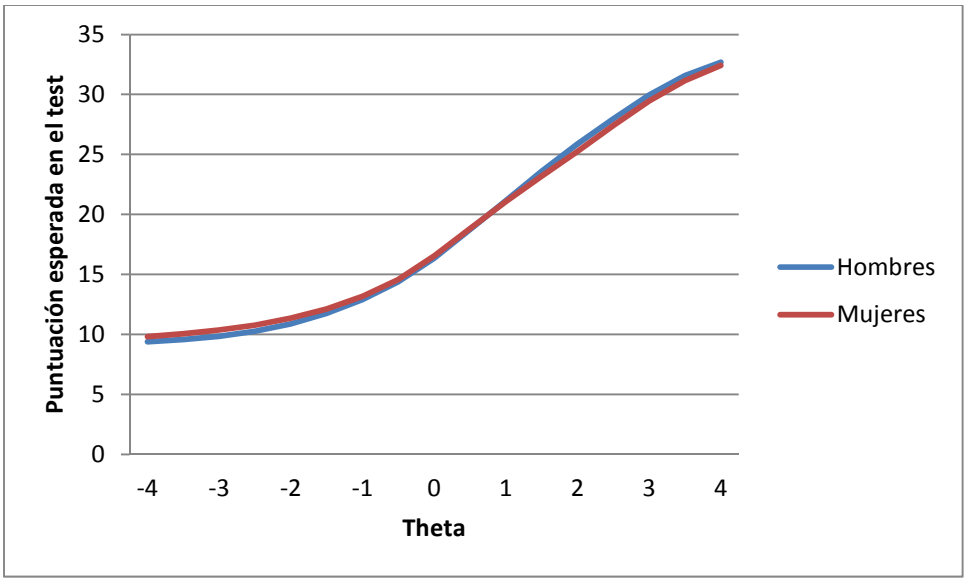



Figura 3.24. CCT para hombres y mujeres en la escala Impulso No Planificado.

### *3.4.1.3. Subescala Impulso Cognitivo-Atencional del BIS*

En primer lugar se pone a prueba la igualdad del parámetro de discriminación entre hombres y mujeres, encontrando que, dado que el valor de $G^2$ es mayor que el punto de corte establecido con la distribución $\chi^2$, se consideran las diferencias entre ambos modelos significativas ($\Delta G^2$ [10] = 26.3, p<.01). El ajuste no es igual de bueno cuando se restringe la igualdad de parámetros *a* en comparación al modelo sin restricciones. No hay equivalencia total de medida, por lo que se analiza el funcionamiento diferencial de todos los ítems de la subescala para buscar la equivalencia en el ámbito de la equivalencia parcial de medida.

En la Tabla 3.37 se indica el DIF relativo a ambos parámetros y, en caso de existir, se desglosa en los parámetros *a* y *b*, junto con el valor del estadístico $G^2$, los grados de libertad correspondientes y el valor estimado para los parámetros en ambos grupos.

Tabla 3.37. *Análisis del funcionamiento diferencial de los ítems de la subescala Impulso Cognitivo-Atencional entre hombres y mujeres*

| | | | | PARÁMETROS HOMBRES | | | | PARÁMETROS MUJERES | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | Hip | $G^2$ | *gl* | *a* | $b_1$ | $b_2$ | $b_3$ | *a* | $b_1$ | $b_2$ | $b_3$ |
| 4 | *ab* igual | 9.1 | 4 | 0.34 | -4.22 | 0.92 | 5.44 | 0.27 | -4.28 | 0.81 | 6.38 |
| 7 | *ab* igual | 11.9 | 4 | 1.55 | -1.18 | -0.19 | 1.45 | 1.13 | -1.42 | -0.08 | 2.08 |
| 7 | *a* igual | 6.0 | 1 | 1.36 | -1.28 | -0.21 | 1.57 | 1.36 | -1.23 | -0.06 | 1.81 |
| 7 | *b* igual | 5.9 | 3 | 1.36 | -1.26 | -0.15 | 1.67 | 1.36 | -1.26 | -0.15 | 1.67 |
| 10 | *ab* igual | 11.8 | 4 | 1.04 | -0.89 | 0.70 | 3.10 | 0.87 | -1.43 | 0.48 | 3.28 |
| 13 | *ab* igual | 16* | 4 | 0.59 | -3.36 | -1.02 | 2.03 | 0.63 | -2.46 | -0.34 | 2.34 |
| 13 | *a* igual | 0.1 | 1 | 0.60 | -3.28 | -0.99 | 1.98 | 0.60 | -2.55 | -0.35 | 2.42 |
| 13 | *b* igual | 15.9* | 3 | 0.59 | -3.02 | -0.72 | 2.23 | 0.59 | -3.02 | -0.72 | 2.23 |
| 16 | *ab* igual | 23.1* | 4 | 1.61 | -0.78 | 0.88 | 1.77 | 1.38 | -0.50 | 0.99 | 1.74 |
| 16 | *a* igual | 1.4 | 1 | 1.51 | -0.81 | 0.91 | 1.84 | 1.51 | -0.47 | 0.93 | 1.64 |
| 16 | *b* igual | 21.7* | 3 | 1.47 | -0.67 | 0.93 | 1.77 | 1.47 | -0.67 | 0.93 | 1.77 |
| 19 | *ab* igual | 5.9 | 4 | 1.16 | -1.23 | 0.19 | 2.36 | 1.42 | -1.22 | 0.18 | 1.99 |
| 20 | *ab* igual | 1.1 | 4 | 1.35 | -0.26 | 1.74 | 2.80 | 1.30 | -0.29 | 1.71 | 2.94 |
| 21 | *ab* igual | 5.9 | 4 | 0.67 | -0.56 | 1.86 | 3.89 | 0.64 | -0.24 | 1.96 | 3.90 |
| 24 | *ab* igual | 15.7* | 4 | 0.98 | 0.09 | 2.16 | 4.03 | 1.01 | -0.12 | 1.62 | 3.17 |
| 24 | *a* igual | 0.0 | 1 | 0.99 | 0.09 | 2.15 | 4.00 | 0.99 | -0.12 | 1.64 | 3.20 |
| 24 | *b* igual | 15.6* | 3 | 0.99 | -0.01 | 1.90 | 3.58 | 0.99 | -0.01 | 1.90 | 3.58 |
| 27 | *ab* igual | 14.9* | 4 | 0.58 | -1.23 | 2.15 | 4.20 | 0.41 | -1.97 | 1.99 | 4.76 |
| 27 | *a* igual | 2.0 | 1 | 0.51 | -1.39 | 2.42 | 4.73 | 0.51 | -1.59 | 1.64 | 3.90 |
| 27 | *b* igual | 13.0* | 3 | 0.51 | -1.48 | 2.04 | 4.32 | 0.51 | -1.48 | 2.04 | 4.32 |

Nota: *$p < 0.01$

Los resultados indican que hay 4 ítems con funcionamiento diferencial: los números 13, 16, 24 y 27. Por otra parte, el ítem que presenta un mayor valor en el incremento de $G^2$ -que roza la significación estadística- respecto al parámetro de discriminación es el ítem 7, por lo que éste será el que se deje variar libremente entre grupos para comparar de nuevo los modelos base y restringido.

Dejando variar libremente el parámetro *a* del ítem 7 entre hombres y mujeres, no hay diferencias significativas entre el modelo base y el modelo con restricciones en cuanto a la discriminación de los ítems en ambos grupos ($\Delta G^2$ [9] = 20.6, n.s.), por lo que se trata de un caso de equivalencia parcial de medida. En la Tabla 3.38 se resumen los cálculos realizados en esta prueba de equivalencia.

Tabla 3.38. *Equivalencia de medida entre hombres y mujeres en la subescala de Impulso Cognitivo-Atencional en relación al parámetro de discriminación*

| | $G^2$ | $\Delta G^2$ | $\Delta$ g.l. | $\chi^2$ | *p* | Ítems libres |
|---|---|---|---|---|---|---|
| Modelo base | 17379.3 | | | | | Todos |
| Modelo invarianza total *a* | 17405.6 | 26.3 | 10 | 23.21 | .01 | Ninguno |
| Modelo invarianza parcial *a* | 17399.9 | 20.6 | 9 | 21.67 | n.s | Ítem 7 |

Para poner a prueba el modelo de invarianza completa de los dos parámetros *a* y *b* entre hombres y mujeres, se compara el modelo base con un modelo que fuerza a ser iguales a ambos parámetros en todos los ítems de la subescala, a excepción del ítem 7, que causó la falta de equivalencia anterior. Los datos confirman la falta de equivalencia ($\Delta G^2$ [36] = 98.4, p<.01), por lo que se eliminan una a una las restricciones de igualdad de los ítems que presentan DIF hasta llegar a la equivalencia parcial (ver Tabla 3.39).

Así, en primer lugar se eliminan las restricciones de igualdad de parámetros del ítem 16, encontrando que las diferencias son significativas ($\Delta G^2$ [32] = 74, p<.01), por lo que se libera además el ítem 13, tras lo cual el $\Delta G^2$ [28] = 54, p<.01. El siguiente ítem del que se eliminan las restricciones de igualdad es el ítem 24, obteniendo un $\Delta G^2$ [24] = 45.6, p<.01). El valor de $\chi^2$ es ligeramente superior al punto de corte, por lo que todavía se hace

necesario liberar un ítem más para encontrar la equivalencia en ambos modelos. Una vez eliminadas las restricciones de igualdad de parámetros también en el ítem 27, no hay diferencias significativas entre ambos modelos ($\Delta G^2$ [20] = 33.2, n.s.), considerándose equivalentes.

Tabla 3.39. *Equivalencia de medida entre hombres y mujeres en la subescala de Impulso Cognitivo-Atencional forzando la igualdad de todos los parámetros*

| | $G^2$ | $\Delta G^2$ | $\Delta$ g.l. | $\chi^2$ | $p$ | Ítems libres |
|---|---|---|---|---|---|---|
| Modelo base | 17379.3 | | | | | Todos |
| Modelo invarianza total *a* y *b* | 17477.7 | 98.4 | 36 | 53.16 | .01 | Ítem 7 |
| Modelo invarianza parcial *a* y *b* | 17453.3 | 74 | 32 | 50.89 | .01 | Ítem 7 y 16 |
| Modelo invarianza parcial *a* y *b* | 17433.3 | 54 | 28 | 48.28 | .01 | Ítem 7, 16 y 13 |
| Modelo invarianza parcial *a* y *b* | 17424.9 | 45.6 | 24 | 42.98 | .01 | Ítem 7, 16, 13 y 24 |
| Modelo invarianza parcial *a* y *b* | 17412.5 | 33.2 | 20 | 37.57 | n.s | Ítem 7, 16, 13, 24 y 27 |

Aunque la mitad de los ítems de la subescala presentan DIF, en la representación gráfica de la CCT para ambos sexos (ver Figura 3.25), ambas líneas están prácticamente solapadas en todo el continuo de $\theta$.

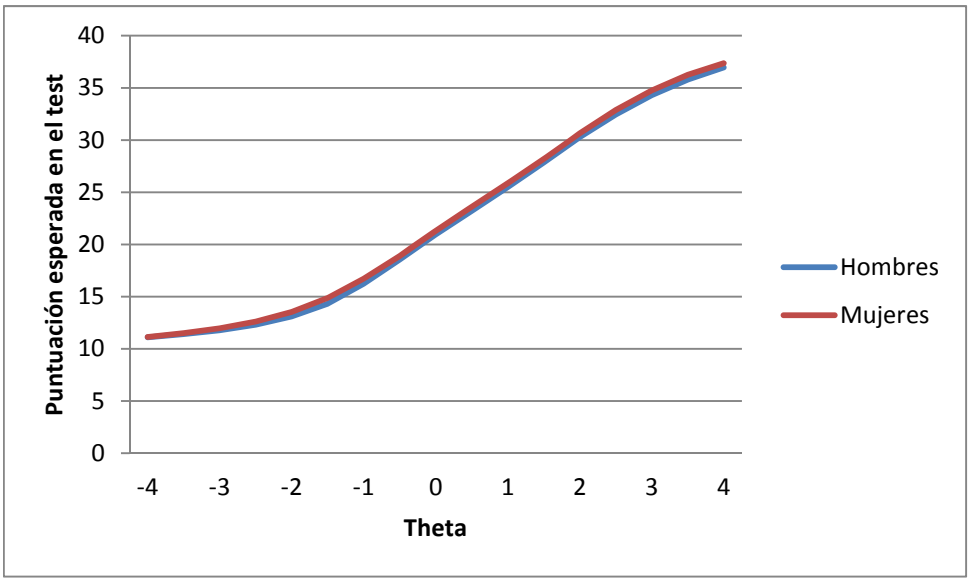



Figura 3.25. CCT para hombres y mujeres en la escala Impulso Cognitivo-Atencional.

### *3.4.1.4. Escala total BIS*

En primer lugar se ponen a prueba la hipótesis de igualad del parámetro de discriminación calculando los valores de verosimilitud del modelo base (sin restricciones de igualdad entre los parámetros de los ítems) y del modelo restringido (que establece la igualdad del parámetro *a* entre hombres y mujeres), obteniendo un $\Delta G^2$ [27] = 81.8, p <.01. Esta diferencia es significativa, por lo que los datos no apoyan la equivalencia de medida entre ambos sexos en la escala BIS completa. Es necesario analizar el funcionamiento diferencial de los ítems para eliminar las restricciones del que contenga una mayor cantidad de DIF y repetir la aplicación del estadístico $G^2$.

En la Tabla 3.40 se muestra el DIF (desglosándose en los dos parámetros en los casos en los que alguno de ellos es significativo o cercano a la significación estadística), junto con el valor del incremento en $G^2$, los grados de libertad y la estimación de los parámetros en ambos grupos.

Tabla 3.40. *Análisis del funcionamiento diferencial de los ítems del test BIS, entre hombres y mujeres*

| | | | | PARÁMETROS HOMBRES | | | | PARÁMETROS MUJERES | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | Hip | $G^2$ | gl | *a* | $b_1$ | $b_2$ | $b_3$ | *a* | $b_1$ | $b_2$ | $b_3$ |
| 1 | *ab* igual | 9.6 | 4 | 1.13 | -1.50 | 0.09 | 3.38 | 1.04 | -1.35 | 0.20 | 4.29 |
| 2 | *ab* igual | 1.0 | 4 | 1.59 | -0.74 | 1.44 | 2.92 | 1.59 | -0.74 | 1.35 | 2.96 |
| 3 | *ab* igual | 16.6 * | 4 | 1.16 | -0.88 | 1.41 | 3.01 | 1.25 | -0.53 | 1.70 | 3.06 |
| 3 | *a* igual | 0.5 | 1 | 1.21 | -0.85 | 1.37 | 2.91 | 1.21 | -0.54 | 1.74 | 3.13 |
| 3 | *b* igual | 16.1* | 3 | 1.22 | -0.67 | 1.55 | 3.01 | 1.22 | -0.67 | 1.55 | 3.01 |
| 4 | *ab* igual | 10.4 | 4 | 0.48 | -2.36 | 0.57 | 3.77 | 0.52 | -2.85 | 0.59 | 3.62 |
| 5 | *ab* igual | 8.2 | 4 | 0.59 | -2.74 | -0.30 | 2.59 | 0.41 | -3.83 | -0.37 | 4.36 |
| 6 | *ab* igual | 6.2 | 4 | 0.45 | 1.31 | 3.27 | 4.37 | 0.45 | 0.90 | 3.35 | 4.48 |
| 7 | *ab* igual | 13.0 | 4 | 1.05 | -1.43 | -0.03 | 2.23 | 1.30 | -1.31 | -0.23 | 1.60 |
| 7 | *a* igual | 3.7 | 1 | 1.19 | -1.29 | -0.01 | 2.04 | 1.19 | -1.40 | -0.24 | 1.71 |
| 7 | *b* igual | 9.3 | 3 | 1.17 | -1.36 | -0.14 | 1.87 | 1.17 | -1.36 | -0.14 | 1.87 |
| 8 | *ab* igual | 14.3* | 4 | 0.56 | -1.57 | 0.67 | 2.87 | 0.68 | -1.64 | 0.02 | 2.10 |
| 8 | *a* igual | 1.2 | 1 | 0.63 | -1.39 | 0.62 | 2.59 | 0.63 | -1.76 | 0.03 | 2.25 |
| 8 | *b* igual | 13.1* | 3 | 0.61 | -1.63 | 0.30 | 2.46 | 0.61 | -1.63 | 0.30 | 2.46 |
| 9 | *ab* igual | 3.1 | 4 | 0.82 | -0.81 | 0.97 | 2.32 | 0.74 | -0.78 | 1.11 | 2.37 |
| 10 | *ab* igual | 5.0 | 4 | 0.91 | -1.27 | 0.56 | 3.23 | 0.94 | -0.98 | 0.73 | 3.33 |
| 11 | *ab* igual | 11.2 | 4 | 1.06 | 0.13 | 1.51 | 3.31 | 1.14 | 0.44 | 1.64 | 3.52 |
| 12 | *ab* igual | 3.4 | 4 | 1.60 | -0.67 | 1.16 | 2.36 | 1.74 | -0.74 | 1.07 | 2.31 |
| 13 | *ab* igual | 14.8* | 4 | 0.37 | -4.00 | -0.54 | 3.83 | 0.34 | -5.49 | -1.64 | 3.33 |
| 13 | *a* igual | 0.1 | 1 | 0.36 | -4.16 | -0.57 | 3.96 | 0.36 | -5.33 | -1.59 | 3.23 |
| 13 | *b* igual | 14.7* | 3 | 0.34 | -5.03 | -1.19 | 3.74 | 0.34 | -5.03 | -1.19 | 3.74 |
| 14 | *ab* igual | 11.7 | 4 | 1.14 | 0.21 | 1.46 | 2.56 | 1.11 | 0.21 | 1.88 | 3.00 |
| 14 | *a* igual | 0.1 | 1 | 1.12 | 0.21 | 1.48 | 2.60 | 1.12 | 0.21 | 1.86 | 2.97 |
| 14 | *b* igual | 11.6 | 3 | 1.12 | 0.21 | 1.67 | 2.78 | 1.12 | 0.21 | 1.67 | 2.78 |
| 15 | *ab* igual | 6.7 | 4 | 1.22 | -0.50 | 1.27 | 2.55 | 1.38 | -0.42 | 0.99 | 2.14 |
| 16 | *ab* igual | 21.1* | 4 | 1.09 | -0.56 | 1.18 | 2.08 | 1.15 | -0.96 | 1.06 | 2.16 |
| 16 | *a* igual | 0.2 | 1 | 1.12 | -0.54 | 1.16 | 2.03 | 1.12 | -0.97 | 1.08 | 2.20 |
| 16 | *b* igual | 20.8* | 3 | 1.10 | -0.79 | 1.13 | 2.15 | 1.10 | -0.79 | 1.13 | 2.15 |
| 17 | *ab* igual | 30.4* | 4 | 0.59 | -0.71 | 0.42 | 2.12 | 0.29 | -1.31 | 1.46 | 5.99 |
| 17 | *a* igual | 7.5* | 1 | 0.41 | -1.06 | 0.53 | 2.91 | 0.41 | -0.95 | 1.05 | 4.32 |
| 17 | *b* igual | 22.8* | 3 | 0.43 | -0.97 | 0.79 | 3.51 | 0.43 | -0.97 | 0.79 | 3.51 |
| 18 | *ab* igual | 7.0 | 4 | 1.86 | -0.35 | 1.26 | 2.37 | 1.79 | -0.45 | 1.45 | 2.41 |
| 19 | *ab* igual | 3.9 | 4 | 1.49 | -1.09 | 0.26 | 2.02 | 1.33 | -1.14 | 0.17 | 2.15 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 20 | *ab* | igual | 1.2 | 4 | 1.23 | -0.22 | 1.83 | 3.10 | 1.25 | -0.28 | 1.82 | 2.93 |
| 21 | *ab* | igual | 9.8 | 4 | 0.78 | -0.12 | 1.75 | 3.37 | 0.84 | -0.47 | 1.54 | 3.21 |
| 22 | *ab* | igual | 11.8 | 4 | 1.17 | -0.14 | 1.34 | 3.24 | 1.28 | -0.07 | 1.00 | 3.08 |
| 22 | *a* | igual | 0.7 | 1 | 1.23 | -0.12 | 1.29 | 3.13 | 1.23 | -0.08 | 1.03 | 3.17 |
| 22 | *b* | igual | 11.1 | 3 | 1.22 | -0.10 | 1.15 | 3.15 | 1.22 | -0.10 | 1.15 | 3.15 |
| 24 | *ab* | igual | 11.5 | 4 | 1.01 | -0.04 | 1.70 | 3.25 | 1.00 | 0.08 | 2.13 | 3.98 |
| 25 | *ab* | igual | 21.5* | 4 | 0.45 | 1.97 | 4.58 | 6.75 | 0.86 | 1.58 | 3.31 | 4.54 |
| 25 | *a* | igual | 8.5* | 1 | 0.66 | 1.45 | 3.26 | 4.76 | 0.66 | 1.95 | 4.13 | 5.70 |
| 25 | *b* | igual | 13.0* | 3 | 0.67 | 1.70 | 3.66 | 5.17 | 0.67 | 1.70 | 3.66 | 5.17 |
| 26 | *ab* | igual | 15.2* | 4 | 0.19 | -0.92 | 7.80 | 14.00 | 0.44 | -0.62 | 3.23 | 5.02 |
| 26 | *a* | igual | 5.6 | 1 | 0.34 | -0.45 | 4.48 | 7.97 | 0.34 | -0.79 | 4.18 | 6.50 |
| 26 | *b* | igual | 9.6 | 3 | 0.33 | -0.65 | 4.37 | 7.20 | 0.33 | -0.65 | 4.37 | 7.20 |
| 27 | *ab* | igual | 10.7 | 4 | 0.67 | -1.17 | 1.39 | 3.17 | 0.74 | -1.02 | 1.75 | 3.41 |
| 29 | *ab* | igual | 23.8* | 4 | 0.49 | -0.16 | 2.93 | 5.50 | 0.62 | -0.89 | 2.09 | 3.93 |
| 29 | *a* | igual | 1.4 | 1 | 0.57 | -0.12 | 2.57 | 4.79 | 0.57 | -0.96 | 2.27 | 4.26 |
| 29 | *b* | igual | 22.4* | 3 | 0.54 | -0.60 | 2.50 | 4.68 | 0.54 | -0.60 | 2.50 | 4.68 |

Nota: *$p < 0.01$

Hay 8 ítems con funcionamiento diferencial en algún parámetro: los números 3, 8, 13, 16, 17, 25 y 29. La mayoría de los ítems presentan funcionamiento diferencial debido a los parámetros *b*, aunque en el caso de los ítems 17 y 25 también existe DIF en relación a la discriminación del ítem.

Aunque el ítem 17 es el que presenta mayor DIF, el ítem 25 es el de mayor desajuste en el parámetro *a* de ambos grupos, por lo que se eliminan las restricciones de igualdad de entre sexos en sus parámetros, volviendo a realizar la comparación entre modelos.

Las diferencias entre ambos modelos siguen siendo significativas ($\Delta G^2$ [26] = 71.5, $p < .01$) por lo que se eliminan, además, las restricciones de igualdad de parámetros del ítem 17. No hay equivalencia entre los modelos ($\Delta G^2$ [25] = 59.7, $p < .01$), por lo que se eliminan las restricciones del ítem 26. De nuevo hay falta de equivalencia entre el modelo

sin restricciones y el modelo restringido ($\Delta G^2$ [24] = 49, p < .01), por lo que se eliminan las restricciones del ítem 29. No hay equivalencia ($\Delta G^2$ [23] = 43.6, p < .01), por lo que ahora se elimina la restricción de igualdad de discriminación entre grupos del ítem 8. El incremento en $G^2$ es ligeramente superior al punto de corte ($\Delta G^2$ [22] = 40.4, p < .01), por lo que todavía hay que eliminar las restricciones de igualdad del parámetro *a* en un ítem más, antes de comparar nuevamente los modelos. Aunque el ítem 3 sería el siguiente con mayor cantidad de DIF, se elimina la restricción de igualdad del parámetro *a* en el ítem 7 porque presenta un mayor DIF no uniforme. Ahora no hay diferencias significativas entre los modelos, ($\Delta G^2$ [21] = 37, ns), con lo que ha sido necesario eliminar 6 ítems para encontrar la equivalencia parcial de medida entre hombres y mujeres en los ítems del test BIS.

En la Tabla 3.41 se resumen todos los datos del proceso de modelos anidado realizado para detectar la posible equivalencia parcial de medida entre ambos grupos.

Tabla 3.41. *Equivalencia de medida entre hombres y mujeres en la escala completa BIS en relación con el parámetro de discriminación*

| | $G^2$ | $\Delta G^2$ | $\Delta$ g.l. | $\chi^2$ | *p* | Ítems libres |
|---|---|---|---|---|---|---|
| Modelo base | 77400.6 | | | | | Todos |
| Modelo invarianza total *a* | 77482.4 | 81.8 | 27 | 46.96 | .01 | Ninguno |
| Modelo invarianza parcial *a* | 77472.1 | 71.5 | 26 | 45.64 | .01 | Ítem 25 |
| Modelo invarianza parcial *a* | 77460.3 | 59.7 | 25 | 44.31 | .01 | Ítem 25 y 17 |
| Modelo invarianza parcial *a* | 77449.6 | 49 | 24 | 42.98 | .01 | Ítem 25, 17 y 26 |
| Modelo invarianza parcial *a* | 77444.2 | 43.6 | 23 | 41.64 | .01 | Ítem 25, 17, 26 y 29 |
| Modelo invarianza parcial *a* | 77441 | 40.4 | 22 | 40.29 | .01 | Item 25, 17, 26, 29 y 8 |
| Modelo invarianza parcial *a* | 77437.6 | 37 | 21 | 38.93 | ns | Ítem 25, 17, 26, 29, 8 y 7 |

Ahora se pone a prueba el modelo más restringido de igualdad de todos los parámetros, $a$, $b_1$, $b_2$ y $b_3$ entre hombres y mujeres en la escala. Para ello, se compara en primer lugar el modelo base, con un modelo que fuerza la igualdad de todos los parámetros en todos los ítems de la subescala, a excepción de los ítems que causaron la falta de equivalencia por DIF no uniforme, es decir, los ítems 25, 17, 26, 29, 8 y 7.

Se concluye que no hay equivalencia total de medida en la escala poniendo a prueba la igualdad de todos los parámetros de los ítems entre hombres y mujeres, puesto que el incremento de $G^2$ es significativo ($\Delta G^2$ [84] = 214, p<.01).

Para comprobar si existe equivalencia parcial se eliminan las restricciones de igualdad de parámetros del ítem 16, por ser el que presenta mayor DIF (ver Tabla 3.40), para comparar de nuevo ambos modelos. Ahora $\Delta G^2$ [80] = 192.4, p<.01, por lo que sigue sin haber equivalencia entre ambos modelos. El siguiente ítem del que se eliminan sus restricciones de igualdad de parámetros es el ítem 3, tras lo cual, el $\Delta G^2$ [76] = 177.8, p<.01. Al no existir igualdad entre los modelos comparados se elimina la restricción de igualdad del último ítem con DIF, el ítem 13, encontrando que las diferencias entre los modelos base y con restricciones siguen siendo significativas ($\Delta G^2$ [72] = 163.4, p<.01).

A pesar de haber liberado de restricciones a todos los ítems que presentan DIF (ver Tabla 3.40), no hay equivalencia parcial de medida entre hombres y mujeres en el global de la escala. Se eliminan ahora las restricciones de los ítems con mayor valor en el incremento de $G^2$ aunque no sean significativos estadísticamente. Así, se libera de la igualdad de parámetros al ítem 22, constatando que las diferencias entre el modelo base y restringido siguen siendo significativas ($\Delta G^2$ [68] = 149, p<.01). El siguiente ítem con

mayor valor en el incremento de $G^2$ es el ítem 14; tras eliminar las restricciones de igualdad de parámetros, hay diferencias significativas ($\Delta G^2$ [64] = 139.3, p<.01). Además, la diferencia entre estos dos últimos modelos no es significativa ($\Delta G^2$ [4] = 9.7, ns) por lo que se para aquí el proceso de modelos anidados, concluyendo que no hay equivalencia parcial en la escala completa BIS en el caso más restrictivo de igualdad de todos los parámetros.

En la Tabla 3.42. se muestran los resultados de poner a prueba la invarianza total de los parámetros de los ítems entre los grupos de hombres y mujeres en la escala completa.

Tabla 3.42. *Equivalencia de medida entre hombres y mujeres en la escala completa BIS forzando la igualdad de todos los parámetros*

| | $G^2$ | $\Delta G^2$ | Δ g.l. | $\chi^2$ | *p* | Ítems libres |
|---|---|---|---|---|---|---|
| Modelo base | 77400.6 | | | | | Todos |
| Modelo invarianza total *a* y *b* | 77614.6 | 214 | 84 | 118.30 | .01 | Ninguno* |
| Modelo invarianza parcial *a* y *b* | 77593 | 192.4 | 80 | 112.33 | .01 | Ítem 16 |
| Modelo invarianza parcial *a* y *b* | 77578.4 | 177.8 | 76 | 112.33 | .01 | Ítem 16 y 3 |
| Modelo invarianza parcial *a* y *b* | 77563.4 | 163.4 | 72 | 100.42 | .01 | Ítem 16, 3 y 13 |
| Modelo invarianza parcial *a* y *b* | 77549.6 | 149 | 68 | 100.42 | .01 | Ítem 16, 3, 13 y 22 |
| Modelo invarianza parcial *a* y *b* | 77539.9 | 139.3 | 64 | 88.38 | .01 | Ítem 16, 3, 13, 22 y 14 |

Nota: * excepto los ítems que causaron la falta de equivalencia en *a* (25, 17, 26, 29, 8 y 7)

A nivel de ítem, los resultados del test completo difieren ligeramente de los encontrados al analizar cada subescala por separado.

(1) En la subescala Impulso Motor se detecta únicamente un ítem con DIF, el ítem 29, mientras que en el análisis del test completo se detecta, además, el ítem 26.

(2) En la subescala Impulso No Planificado, en ambos casos se detectan 4 ítems, pero hay uno que no es coincidente: el ítem 3 presenta DIF en el análisis de la escala completa pero no en el de las subescalas, y el ítem 22 es el caso contrario, ya que presenta DIF en el análisis de la subescala pero no en de la escala completa.

(3) En la subescala Impulso Cognitivo-Atencional es donde se encuentra más diferencias, habiendo únicamente dos ítems con DIF en el análisis de la escala completa (13 y 16), y 4 en el análisis de la subescala (13, 16, 24 y 27).

A pesar de la falta de equivalencia en la escala BIS encontrada con el procedimiento de comparación de modelos basado en el test LR, en la representación de la CCT para ambos sexos no se aprecian visualmente diferencias entre hombres y mujeres en la puntuación esperada del test (ver Figura 3.26).

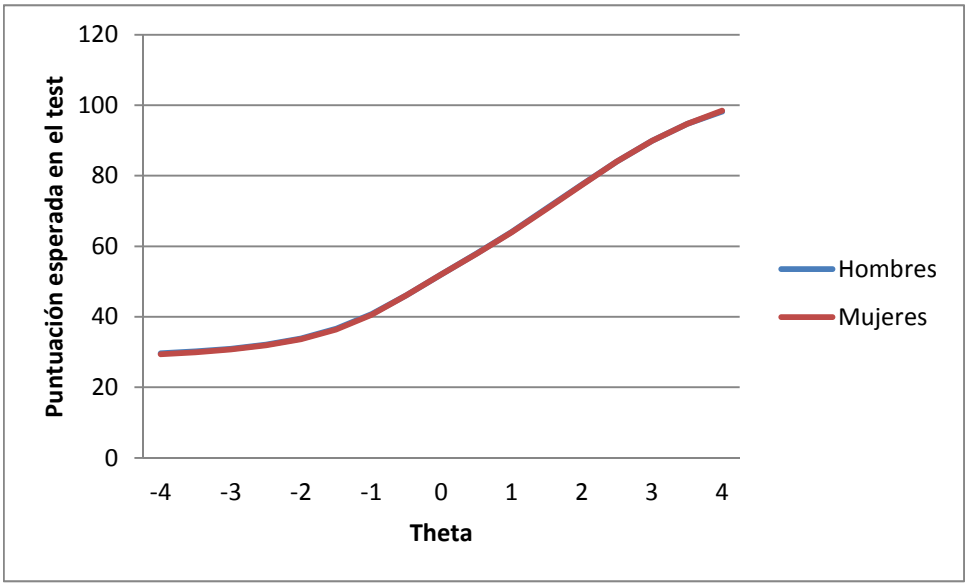



Figura 3.26. CCT para hombres y mujeres en la escala BIS.

## 3.4.2. EQUIVALENCIA DE MEDIDA ENTRE PREADOLESCENTES Y ADOLESCENTES

### *3.4.2.1. Subescala Impulso motor del BIS*

En primer lugar se pone a prueba la igualdad del parámetro de discriminación entre preadolescentes y adolescentes, encontrando que no hay equivalencia total, ya que el incremento de $G^2$ es significativo ($\Delta G^2$ [8] = 38, p<.01). Al no encontrar equivalencia en el total de la subescala se busca qué ítems tienen mayores niveles de DIF para buscar la equivalencia en el marco de la equivalencia parcial de medida. En la Tabla 3.43 aparecen los resultados del análisis DIF entre preadolescentes y adolescentes en la subescala de Impulso Motor. En el caso de los ítems que no presentan DIF se presenta únicamente una línea que corresponde con los resultados de poner a prueba la igualdad de los dos grupos en todos los parámetros del ítem y su estimación correspondiente. En el caso de los ítems que sí presentan DIF, además de esta información se incluyen, en dos líneas más, los resultados del DIF en el parámetro *a* y en los parámetros *b*.

Tabla 3.43. *Análisis del funcionamiento diferencial de los ítems de la subescala Impulso Motor entre preadolescente y adolescentes*

| | | | | PARÁMETROS PREADOLESCENTES | | | | PARÁMETROS ADOLESCENTES | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | Hip | $G^2$ | g.l. | *a* | $b_1$ | $b_2$ | $b_3$ | *a* | $b_1$ | $b_2$ | $b_3$ |
| 2 | *ab* igual | 74.4* | 4 | 1.13 | -0.96 | 1.70 | 3.34 | 2.62 | -1.05 | 0.88 | 2.24 |
| 2 | *a* igual | 46.5* | 1 | 1.82 | -0.84 | 1.12 | 2.27 | 1.82 | -1.24 | 1.01 | 2.63 |
| 2 | *b* igual | 27.9* | 3 | 1.86 | -1.05 | 1.03 | 2.48 | 1.86 | -1.05 | 1.03 | 2.48 |
| 6 | *ab* igual | 5.7 | 4 | 0.35 | 0.71 | 3.54 | 4.83 | 0.41 | 1.08 | 3.53 | 4.85 |
| 9 | *ab* igual | 7.1 | 4 | 0.65 | -1.41 | 0.92 | 2.57 | 0.82 | -0.96 | 0.73 | 1.93 |
| 12 | *ab* igual | 14.2* | 4 | 1.92 | -0.99 | 0.71 | 1.72 | 2.47 | -0.97 | 0.73 | 1.86 |
| 12 | *a* igual | 4.1 | 1 | 2.24 | -0.96 | 0.62 | 1.55 | 2.24 | -1.01 | 0.76 | 1.94 |
| 12 | *b* igual | 10.1 | 3 | 2.18 | -0.99 | 0.73 | 1.85 | 2.18 | -0.99 | 0.73 | 1.85 |
| 15 | *ab* igual | 0.8 | 4 | 1.19 | -0.79 | 0.94 | 2.21 | 1.27 | -0.82 | 0.83 | 2.12 |
| 18 | *ab* igual | 47.4 * | 4 | 1.77 | -0.71 | 0.95 | 1.70 | 2.60 | -0.71 | 1.00 | 2.13 |
| 18 | *a* igual | 9.1* | 1 | 2.22 | -0.69 | 0.79 | 1.44 | 2.22 | -0.76 | 1.06 | 2.26 |
| 18 | *b* igual | 38.3* | 3 | 2.08 | -0.74 | 1.02 | 2.05 | 2.08 | -0.74 | 1.02 | 2.05 |
| 26 | *ab* igual | 8.1 | 4 | 0.43 | -0.82 | 3.62 | 6.21 | 0.42 | -0.86 | 2.93 | 5.05 |
| 29 | *ab* igual | 11.0 | 4 | 0.60 | -1.18 | 1.68 | 3.50 | 0.56 | -0.69 | 2.34 | 4.60 |

Nota: *$p < 0.01$

Hay tres ítems que presentan funcionamiento diferencial. El ítem que presenta un mayor incremento de $G^2$ es el ítem 2 (ver Tabla 3.43) con un valor altísimo, no visto hasta ahora en ninguno de los resultados de la variable sexo ya analizada. Eliminando las restricciones de igualdad de parámetro *a* en este ítem, el $\Delta G^2$ [7] = 8.6, n.s., por lo que hay equivalencia de parcial de medida entre preadolescentes y adolescentes en cuanto al parámetro de discriminación. En la siguiente Tabla se describen estos resultados.

Tabla 3.44. *Equivalencia de medida entre preadolescentes y adolescentes en la subescala de Impulso Motor en relación con el parámetro de discriminación*

| | $G^2$ | $\Delta G^2$ | $\Delta$ g.l. | $\chi^2$ | $p$ | Ítems libres |
|---|---|---|---|---|---|---|
| Modelo base | 6590.2 | | | | | Todos |
| Modelo invarianza total *a* | 6628.1 | 38 | 8 | 20.09 | .01 | Ninguno |
| Modelo invarianza parcial *a* | 6598.8 | 8.6 | 7 | 18.48 | n.s | Ítem 2 |

A continuación se pone a prueba el modelo más restrictivo de igualdad de parámetros, forzando la igualdad de $a$, $b_1$, $b_2$ y $b_3$ entre ambos grupos.

Se concluye que puesto que el incremento de $G^2$ es significativo ($\Delta G^2$ [28] = 81.2, $p<.01$) no hay equivalencia total de medida en la subescala poniendo a prueba la igualdad de todos los parámetros de los ítems -a excepción del ítem 2 cuya igualdad se descartó en el paso anterior-. Para averiguar si la hay en el ámbito de la equivalencia parcial de medida se eliminan las restricciones del ítem con mayor incremento en $G^2$, esto es el ítem 18.

En esta ocasión no hay diferencias entre el modelo base y el modelo con restricciones de igualdad de parámetros $\Delta G^2$ [24] = 41.5, n.s., por lo que hay equivalencia parcial de medida entre preadolescentes y adolescentes en la subescala. Los únicos ítems que rompen la equivalencia total de medida son el ítem 2 y el ítem 18. En la Tabla 3.45 se reproducen estos resultados.

Tabla 3.45. *Equivalencia de medida entre preadolescentes y adolescentes en la subescala de Impulso Motor forzando la igualdad de todos los parámetros*

| | $G^2$ | $\Delta G^2$ | $\Delta$ g.l. | $\chi^2$ | *p* | Ítems libres |
|---|---|---|---|---|---|---|
| Modelo base | 6590.2 | | | | | Todos |
| Modelo invarianza total *a* y *b* | 6671.4 | 81.2 | 28 | 48.28 | .01 | Ítem 2 |
| Modelo invarianza parcial *a* y *b* | 6631.7 | 41.5 | 24 | 42.98 | ns | Ítems 2 y 18 |

Gráficamente, las diferencias en la puntuación esperada de la subescala IM entre preadolescentes y adolescentes son escasas, como se muestran en la Figura 3.27.

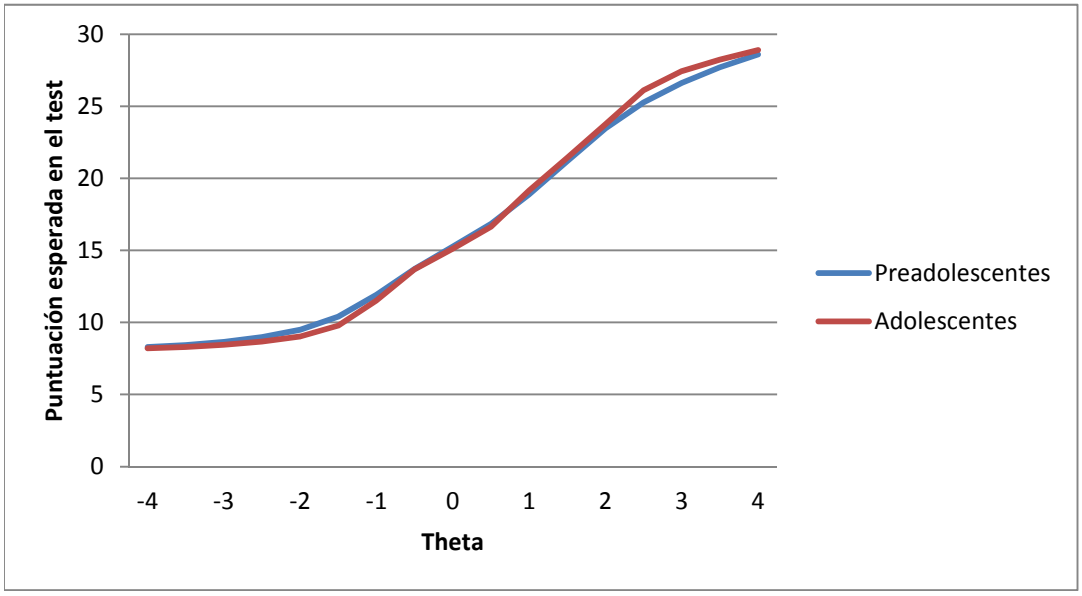



Figura 3.27. CCT para ambos grupos de edad en la subescala Impulso Motor

### *3.4.2.2. Subescala Impulso no Planificado del BIS*

Según los resultados, no hay equivalencia total de medida en relación al parámetro de discriminación entre ambos grupos, ya que el incremento de $G^2$ es significativo ($\Delta G^2$ [9] =

24.7, p<.01). Se analiza el funcionamiento diferencial de todos los ítems de la subescala para buscar cuáles rompen la equivalencia total, en el ámbito de la equivalencia parcial de medida.

En la Tabla 3.46 se presentan los resultados del análisis DIF de los ítems de la subescala INP entre preadolescentes y adolescentes. En los ítems con funcionamiento diferencial, la información se ha desglosado en DIF uniforme, no uniforme, y ambos. En todos los casos se incluye además la correspondiente estimación de los parámetros.

Tabla 3.46. *Análisis del funcionamiento diferencial de los ítems de la subescala Impulso No Planificado entre preadolescentes y adolescentes*

| | | | | PARÁMETROS PREADOLESCENTES | | | | PARÁMETROS ADOLESCENTES | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | Hip | $G^2$ | g.l. | *a* | $b_1$ | $b_2$ | $b_3$ | *a* | $b_1$ | $b_2$ | $b_3$ |
| 1 | *ab* igual | 25.9* | 4 | 0.70 | -2.17 | -0.16 | 4.72 | 1.14 | -2.03 | -0.24 | 3.51 |
| 1 | *a* igual | 9.3* | 1 | 0.93 | -1.91 | -0.31 | 3.49 | 0.93 | -2.38 | -0.28 | 4.14 |
| 1 | *b* igual | 16.7* | 3 | 0.95 | -2.09 | -0.28 | 3.85 | 0.95 | -2.09 | -0.28 | 3.85 |
| 3 | *ab* igual | 53.9* | 4 | 0.73 | -1.88 | 1.65 | 3.45 | 1.35 | -0.80 | 1.26 | 2.82 |
| 3 | *a* igual | 16.1* | 1 | 1.06 | -1.64 | 1.00 | 2.31 | 1.06 | -0.95 | 1.48 | 3.36 |
| 3 | *b* igual | 37.8* | 3 | 0.85 | -1.39 | 1.61 | 3.60 | 0.85 | -1.39 | 1.61 | 3.60 |
| 5 | *ab* igual | 27.9* | 4 | 0.28 | -5.34 | -1.09 | 4.51 | 0.53 | -3.71 | -0.60 | 3.16 |
| 5 | *a* igual | 5.2 | 1 | 0.40 | -4.03 | -1.04 | 2.90 | 0.40 | -4.79 | -0.77 | 4.07 |
| 5 | *b* igual | 22.6* | 3 | 0.38 | -4.63 | -0.89 | 3.85 | 0.38 | -4.63 | -0.89 | 3.85 |
| 8 | *ab* igual | 34.3* | 4 | 0.28 | -2.58 | 1.13 | 5.86 | 0.61 | -2.53 | -0.30 | 1.90 |
| 8 | *a* igual | 8.4* | 1 | 0.45 | -1.97 | 0.42 | 3.43 | 0.45 | -3.32 | -0.39 | 2.50 |
| 8 | *b* igual | 25.9* | 3 | 0.57 | -2.25 | -0.11 | 2.25 | 0.57 | -2.25 | -0.11 | 2.25 |
| 11 | *ab* igual | 42.5* | 4 | 1.05 | 0.31 | 1.55 | 2.85 | 1.49 | -0.30 | 0.94 | 2.71 |
| 11 | *a* igual | 4.8 | 1 | 1.32 | 0.16 | 1.21 | 2.30 | 1.32 | -0.32 | 1.01 | 2.93 |
| 11 | *b* igual | 37.6* | 3 | 1.42 | -0.16 | 1.01 | 2.59 | 1.42 | -0.16 | 1.01 | 2.59 |
| 14 | *ab* igual | 32.9* | 4 | 1.11 | -0.24 | 1.21 | 1.84 | 1.25 | -0.22 | 1.26 | 2.61 |
| 14 | *a* igual | 0.5 | 1 | 1.19 | -0.26 | 1.12 | 1.72 | 1.19 | -0.22 | 1.30 | 2.70 |
| 14 | *b* igual | 32.4* | 3 | 1.13 | -0.23 | 1.31 | 2.45 | 1.13 | -0.23 | 1.31 | 2.45 |
| 17 | *ab* igual | 16.8* | 4 | 0.44 | -0.95 | 0.94 | 3.61 | 0.41 | -1.86 | -0.08 | 2.82 |
| 17 | *a* igual | 0.1 | 1 | 0.42 | -0.95 | 1.01 | 3.78 | 0.42 | -1.82 | -0.08 | 2.75 |
| 17 | *b* igual | 16.8* | 3 | 0.52 | -1.26 | 0.23 | 2.51 | 0.52 | -1.26 | 0.23 | 2.51 |
| 22 | *ab* igual | 63.5* | 4 | 1.41 | -0.55 | 0.34 | 1.77 | 1.77 | -0.50 | 0.76 | 2.63 |
| 22 | *a* igual | 1.9 | 1 | 1.63 | -0.57 | 0.23 | 1.53 | 1.63 | -0.53 | 0.79 | 2.75 |
| 22 | *b* igual | 61.6* | 3 | 1.40 | -0.53 | 0.71 | 2.58 | 1.40 | -0.53 | 0.71 | 2.58 |
| 25 | *ab* igual | 12.7 | 4 | 0.53 | 1.69 | 3.83 | 5.07 | 0.47 | 1.82 | 4.73 | 7.43 |

Nota: *$p < 0.01$

Todos los ítems de la escala, a excepción del ítem 25, presentan funcionamiento diferencial (ver Tabla 3.46). De ellos, el que presenta un mayor incremento de $G^2$ respecto al parámetro de discriminación es el ítem 3. Por este motivo, se libera de restricciones a este ítem y se comparan de nuevo los modelos libre y con restricciones de igualdad en el parámetro *a*, obteniendo un $\Delta G^2$ [8] = 19.4, n.s., por lo que la subescala presenta

equivalencia parcial de medida entre preadolescentes y adolescentes en cuanto al parámetro de discriminación. El ítem 3 es el único que rompe la equivalencia total de medida. En la Tabla 3.47 aparecen desglosados estos resultados.

Tabla 3.47. *Equivalencia de medida entre preadolescentes y adolescentes en la subescala de Impulso No Planificado en relación al parámetro de discriminación*

| | $G^2$ | $\Delta G^2$ | $\Delta$ g.l. | $\chi^2$ | $p$ | Ítems libres |
|---|---|---|---|---|---|---|
| Modelo base | 10690.2 | | | | | Todos |
| Modelo invarianza total *a* | 10714.9 | 24.7 | 9 | 21.67 | .01 | Ninguno |
| Modelo invarianza parcial *a* | 10709.6 | 19.4 | 8 | 20.1 | ns | Ítem 3 |

Para poner a prueba la condición más restrictiva de igualdad de parámetros entre preadolescentes y adolescentes, se fuerza la igualdad entre ambos grupos en todos los parámetros ($a$, $b_1$, $b_2$ y $b_3$) de los ítems de la subescala (a excepción del ítem 3), comparando este modelo restringido con el modelo base. Los resultados indican una clara falta de equivalencia ($\Delta G^2$ [32] = 242.2, p<.01).

A continuación se utiliza la Tabla 3.46 para eliminar las restricciones del ítem con mayor incremento en $G^2$, antes de volver a comparar el modelo sin restricciones con el modelo base. De esta manera, se eliminan sucesivamente las restricciones de los ítems 22, 11, 8, 14, 5, 1 y 17 (ver Tabla 3.46 para consultar los datos de los índices calculados) llegando a un modelo con restricciones que sigue siendo significativamente diferente del modelo base ($\Delta G^2$ [4] = 13.8, p<.01), a pesar de que sólo uno de los ítems se ha forzado a tener el mismo valor para ambos parámetros en los dos grupos.

Por tanto, en la subescala de Impulso No planificado no hay equivalencia de medida con respecto a la edad, ni total, ni parcial, por lo que no sería apropiada su aplicación para preadolescentes y adolescentes indistintamente.

Tabla 3.48. *Equivalencia de medida entre preadolescentes y adolescentes en la subescala de Impulso No Planificado forzando la igualdad de todos los parámetros*

| | $G^2$ | $\Delta G^2$ | $\Delta$ g.l. | $\chi^2$ | $p$ | Ítems libres |
|---|---|---|---|---|---|---|
| Modelo base | 10690.2 | | | | | Todos |
| Modelo invarianza total *a* y *b* | 10932.4 | 242.2 | 32 | 43.8 | 0.01 | Ítem 3 |
| Modelo invarianza parcial *a* y *b* | 10862.2 | 172 | 28 | 48.3 | 0.01 | Ítem 3 y 22 |
| Modelo invarianza parcial *a* y *b* | 10832.1 | 141.9 | 24 | 43 | 0.01 | Ítem 3, 22 y 11 |
| Modelo invarianza parcial *a* y *b* | 10806.7 | 116.5 | 20 | 37.6 | 0.01 | Ítem 3, 22, 11 y 8 |
| Modelo invarianza parcial *a* y *b* | 10767.2 | 77 | 16 | 32 | 0.01 | Ítem 3, 22, 11, 8 y 14 |
| Modelo invarianza parcial *a* y *b* | 10739.3 | 49.1 | 12 | 26.2 | 0.01 | Ítem 3, 22, 11, 8, 14 y 5 |
| Modelo invarianza parcial *a* y *b* | 10714.5 | 24.3 | 8 | 20.1 | 0.01 | Ítem 3, 22, 11, 8, 14, 5 y 1 |
| Modelo invarianza parcial *a* y *b* | 10704 | 13.8 | 4 | 13.3 | 0.01 | Ítem 3, 22, 11, 8, 14, 5, 1 y 17 |

Las diferencias en puntuación esperada en la subescala INP se muestran en la Figura 3.28. A niveles bajos del rasgo la probabilidad de obtener una puntuación más alta es mayor para los preadolescentes, invirtiéndose esta relación en los niveles altos del rasgo, en los que la probabilidad de obtener una puntuación más alta es para los adolescentes.

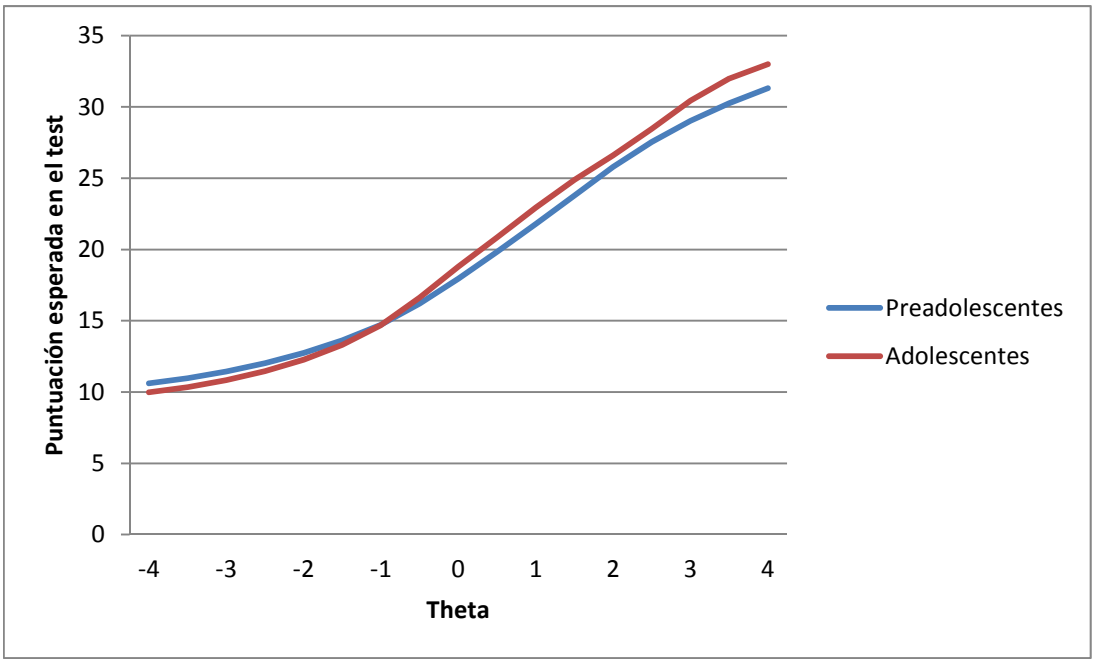



Figura 3.28. CCT para ambos grupos de edad en la subescala Impulso No Planificado.

### *3.4.2.3. Subescala Impulso Cognitivo-Atencional del BIS*

Al poner a prueba la igualdad del parámetro *a* entre ambos grupos de edad se concluye que no hay equivalencia total de medida, ya que el incremento en $G^2$ es significativo ($\Delta G^2$ [10] = 26.9, $p$<.01).

Para averiguar qué ítems rompen la equivalencia de medida se analiza el funcionamiento diferencial de todos los ítems de la subescala, encontrando que la mayoría de ellos presenta DIF (ver Tabla 3.49, en la que se presentan, en caso de ser significativos, los resultados desglosados por parámetros).

Tabla 3.49. *Análisis del funcionamiento diferencial de los ítems de la escala Impulso Cognitivo-Atencional entre preadolescentes y adolescentes*

| | | | | PARÁMETROS PREADOLESCENTES | | | | PARÁMETROS ADOLESCENTES | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Ítem | Hip | $G^2$ | g.l. | *a* | $b_1$ | $b_2$ | $b_3$ | *a* | $b_1$ | $b_2$ | $b_3$ |
| 4 | *ab* igual | 83.4* | 4 | 0.26 | -7.35 | -0.36 | 6.54 | 0.10 | -7.66 | 6.40 | 17.27 |
| 4 | *a* igual | 2.2 | 1 | 0.17 | -10.83 | -0.53 | 9.63 | 0.17 | -4.94 | 3.61 | 10.21 |
| 4 | *b* igual | 81.3* | 3 | 0.31 | -4.58 | 0.55 | 5.48 | 0.31 | -4.58 | 0.55 | 5.48 |
| 7 | *ab* igual | 13.2 | 4 | 1.32 | -1.75 | -0.51 | 1.42 | 0.89 | -1.93 | -0.38 | 2.31 |
| 7 | *a* igual | 8.4 | 1 | 1.10 | -1.99 | -0.57 | 1.61 | 1.10 | -1.75 | -0.43 | 1.88 |
| 7 | *b* igual | 5.2 | 3 | 1.15 | -1.83 | -0.51 | 1.63 | 1.15 | -1.83 | -0.51 | 1.63 |
| 10 | *ab* igual | 68.7* | 4 | 0.98 | -1.19 | 0.69 | 3.34 | 0.91 | -1.95 | -0.33 | 2.40 |
| 10 | *a* igual | 0.2 | 1 | 0.95 | -1.23 | 0.71 | 3.43 | 0.95 | -1.92 | -0.34 | 2.31 |
| 10 | *b* igual | 68.5* | 3 | 0.74 | -1.74 | 0.43 | 3.69 | 0.74 | -1.74 | 0.43 | 3.69 |
| 13 | *ab* igual | 73.8* | 4 | 0.55 | -4.13 | -0.92 | 2.68 | 0.51 | -3.33 | -1.48 | 1.35 |
| 13 | *a* igual | 0.1 | 1 | 0.53 | -4.25 | -0.95 | 2.77 | 0.53 | -3.26 | -1.46 | 1.29 |
| 13 | *b* igual | 73.7* | 3 | 0.48 | -4.00 | -1.21 | 2.37 | 0.48 | -4.00 | -1.21 | 2.37 |
| 16 | *ab* igual | 16.4* | 4 | 1.37 | -1.11 | 0.80 | 1.79 | 1.21 | -1.09 | 0.59 | 1.47 |
| 16 | *a* igual | 0.9 | 1 | 1.29 | -1.15 | 0.83 | 1.86 | 1.29 | -1.08 | 0.54 | 1.38 |
| 16 | *b* igual | 15.5* | 3 | 1.27 | -1.12 | 0.75 | 1.72 | 1.27 | -1.12 | 0.75 | 1.72 |
| 19 | *ab* igual | 10.0 | 4 | 1.11 | -1.67 | -0.00 | 2.32 | 1.04 | -1.93 | -0.31 | 2.02 |
| 20 | *ab* igual | 17.4* | 4 | 1.33 | -0.69 | 1.52 | 2.79 | 0.90 | -0.47 | 2.07 | 3.39 |
| 20 | *a* igual | 7.3* | 1 | 1.12 | -0.78 | 1.71 | 3.16 | 1.12 | -0.50 | 1.67 | 2.79 |
| 20 | *b* igual | 10.1 | 3 | 1.17 | -0.65 | 1.65 | 2.94 | 1.17 | -0.65 | 1.65 | 2.94 |
| 21 | *ab* igual | 40.4* | 4 | 0.51 | -1.42 | 1.70 | 4.55 | 0.48 | -0.13 | 2.92 | 5.00 |
| 21 | *a* igual | 0.2 | 1 | 0.50 | -1.46 | 1.75 | 4.67 | 0.50 | -0.15 | 2.81 | 4.82 |
| 21 | *b* igual | 40.2* | 3 | 0.60 | -0.80 | 1.76 | 3.95 | 0.60 | -0.80 | 1.76 | 3.95 |
| 24 | *ab* igual | 32.4* | 4 | 0.90 | -0.51 | 1.81 | 4.08 | 0.74 | -0.02 | 2.09 | 3.63 |
| 24 | *a* igual | 1.4 | 1 | 0.83 | -0.55 | 1.95 | 4.41 | 0.83 | -0.07 | 1.87 | 3.28 |
| 24 | *b* igual | 31.0* | 3 | 0.87 | -0.34 | 1.83 | 3.73 | 0.87 | -0.34 | 1.83 | 3.73 |
| 27 | *ab* igual | 12.5 | 4 | 0.57 | -1.72 | 1.55 | 3.85 | 0.31 | -2.39 | 3.26 | 6.34 |

Nota: *$p < 0.01$

Se decide eliminar las restricciones de igualdad del parámetro *a* entre ambos grupos del ítem 4, aunque no es el que concentra mayor DIF en *a*, porque su incremento en $G^2$ es muy elevado.

La comparación del modelo que deja variar libremente el parámetro de discriminación de todos los ítems entre ambos grupos con el modelo con restricciones de igualdad en todos los ítems excepto en el 4 resulta no ser significativa ($\Delta G^2$ [9] = 14.6, n.s.). Por tanto, hay equivalencia parcial de medida entre preadolescentes y adolescentes, existiendo un único ítem que rompe la equivalencia total, el ítem 4.

Tabla 3.50. *Equivalencia de medida entre preadolescentes y adolescentes en la subescala de Impulso No Planificado en relación al parámetro de discriminación*

| | $G^2$ | $\Delta G^2$ | $\Delta$ g.l. | $\chi^2$ | $p$ | Ítems libres |
|---|---|---|---|---|---|---|
| Modelo base | 17000.8 | | | | | Todos |
| Modelo invarianza total *a* | 17027.7 | 26.9 | 10 | 23.21 | .01 | Ninguno |
| Modelo invarianza parcial *a* | 17015.4 | 14.6 | 9 | 21.67 | ns | Ítem 4 |

A continuación se pone a prueba la hipótesis más restrictiva de igualdad de todos los parámetros de los ítems entre los dos grupos de edad, encontrando tras la comparación del modelo base con el modelo de invarianza total que las diferencias son significativas ($\Delta G^2$ [36] = 286.7, $p$<.01).

No hay invarianza total de medida forzando la igualdad de todos los parámetros, por lo que se eliminan una a una las restricciones de igualdad de los ítems para comprobar si existe invarianza parcial. En primer lugar, se liberó el ítem 13, por ser el de mayor desajuste en sus parámetros entre grupos, resultando la comparación de modelos significativa ($\Delta G^2$ [32] = 211.7, $p$<.01). Después, se eliminaron sucesivamente las restricciones de igualdad de parámetros de los ítems 10, 21, 24, 16, 20 y 7, sin conseguir la equivalencia entre el modelo base y el modelo con restricciones ($\Delta G^2$ [8] = 20.8, $p$<.01). Dado que la diferencia entre el último modelo que se comparó con el modelo base y el

anterior no es significativa ($\Delta G^2$ [4] = 10.2, *ns*) se interrumpe el proceso de modelos anidados, concluyendo que no hay equivalencia de medida parcial entre preadolescentes y adolescentes en el modelo más restrictivo de igualdad de todos los parámetros. Estos resultados se detallan en la Tabla 3.51.

Tabla 3.51. *Equivalencia de medida entre preadolescentes y adolescentes en la subescala de Impulso Cognitivo-Atencional forzando la igualdad de todos los parámetros*

| | $G^2$ | $\Delta G^2$ | $\Delta$ g.l. | $\chi^2$ | *p* | Ítems libres |
|---|---|---|---|---|---|---|
| Modelo base | 17000.8 | | | | | Todos |
| Modelo invarianza total *a* y *b* | 17287.5 | 286.7 | 36 | 53.2 | .01 | Ítem 4 |
| Modelo invarianza parcial *a* y *b* | 17212.5 | 211.7 | 32 | 43.8 | .01 | Ítems 4 y 13 |
| Modelo invarianza parcial *a* y *b* | 17141.5 | 140.7 | 28 | 48.3 | .01 | Ítems 4, 13 y 10 |
| Modelo invarianza parcial *a* y *b* | 17104.0 | 103.2 | 24 | 43 | .01 | Ítems 4, 13, 10 y 21 |
| Modelo invarianza parcial *a* y *b* | 17070.0 | 69.2 | 20 | 37.6 | .01 | Ítems 4, 13, 10, 21 y 24 |
| Modelo invarianza parcial *a* y *b* | 17047.6 | 46.8 | 16 | 32 | .01 | Ítems 4, 13, 10, 21, 24 y 16 |
| Modelo invarianza parcial *a* y *b* | 17031.8 | 31 | 12 | 26.2 | .01 | Ítems 4, 13, 10, 21, 24, 16 y 20 |
| Modelo invarianza parcial *a* y *b* | 17021.6 | 20.8 | 8 | 20.1 | .01 | Ítems 4, 13, 10, 21, 24, 16, 20 y 7 |

Gráficamente, las diferencias entre preadolescentes y adolescentes se encuentran en los niveles altos del rasgo, en los que los preadolescentes tienen una mayor puntuación esperada que los adolescentes (ver Figura 3.29).

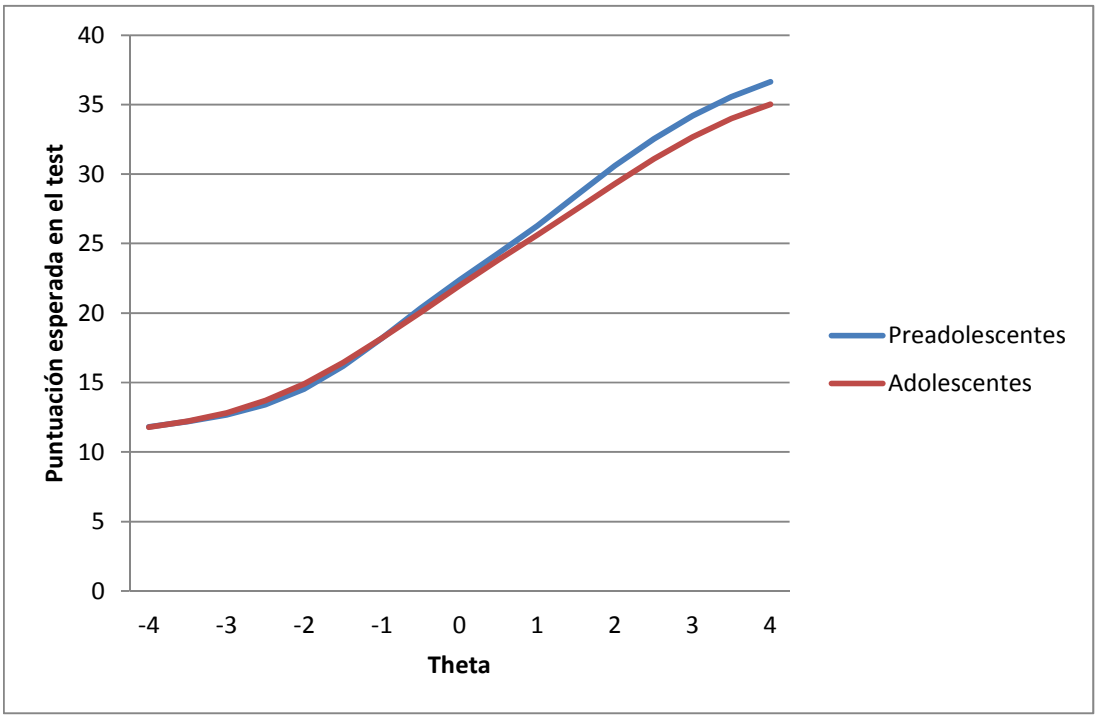



Figura 3.29. CCT para ambos grupos de edad en la subescala Impulso Cognitivo-Atencional.

### *3.4.2.4. Escala total BIS*

Se comprueba, en primer lugar, la igualdad del parámetro de discriminación entre ambos grupos de edad. Para ello, se comparan los valores de verosimilitud del modelo sin restricciones de igualdad entre los parámetros de los grupos con los del modelo que exige la igualdad del parámetro *a* en todos los ítems entre preadolescentes y adolescentes.

El incremento de $G^2$ entre ambos modelos es significativo ($\Delta G^2$ [27] = 133.6, $p$<.01), por tanto, los resultados no apoyan la equivalencia total de medida entre preadolescentes y adolescentes en la escala.

Para comprobar si existe equivalencia parcial de medida entre los grupos se analiza el funcionamiento diferencial de todos los ítems de la escala (ver Tabla 3.52, donde se muestra el incremento en $G^2$ desglosado en función de los parámetros en caso de significación estadística, además de la estimación de los parámetros en cada caso para ambos grupos).

Tabla 3.52. *Análisis del funcionamiento diferencial de los ítems del test BIS entre preadolescentes y adolescentes*

| | | | | PARÁMETROS PREADOLESCENTES | | | | PARÁMETROS ADOLESCENTES | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Ítem | Hip | $G^2$ | g.l. | *a* | $b_1$ | $b_2$ | $b_3$ | *a* | $b_1$ | $b_2$ | $b_3$ |
| 1 | *ab* igual | 35.2* | 4 | 1.14 | -2.03 | -0.25 | 3.54 | 0.70 | -2.01 | -0.04 | 4.78 |
| 1 | *a* igual | 12.9* | 1 | 0.93 | -2.38 | -0.28 | 4.16 | 0.93 | -1.76 | -0.19 | 3.61 |
| 1 | *b* igual | 22.3* | 3 | 0.96 | -1.99 | -0.24 | 3.85 | 0.96 | -1.99 | -0.24 | 3.85 |
| 2 | *ab* igual | 44.6* | 4 | 1.76 | -1.23 | 1.01 | 2.66 | 0.99 | -1.17 | 1.72 | 3.54 |
| 2 | *a* igual | 27.1* | 1 | 1.37 | -1.43 | 1.16 | 3.12 | 1.37 | -1.09 | 1.24 | 2.64 |
| 2 | *b* igual | 17.4* | 3 | 1.42 | -1.24 | 1.16 | 2.90 | 1.42 | -1.24 | 1.16 | 2.90 |
| 3 | *ab* igual | 47.5* | 4 | 1.43 | -0.78 | 1.21 | 2.71 | 0.92 | -1.52 | 1.39 | 2.87 |
| 3 | *a* igual | 14.5* | 1 | 1.18 | -0.88 | 1.37 | 3.10 | 1.18 | -1.39 | 1.02 | 2.23 |
| 3 | *b* igual | 33.0* | 3 | 1.06 | -1.18 | 1.37 | 3.05 | 1.06 | -1.18 | 1.37 | 3.05 |
| 4 | *ab* igual | 64.6* | 4 | 0.40 | -4.75 | -0.26 | 4.19 | 0.30 | -3.28 | 1.73 | 5.61 |
| 4 | *a* igual | 1.2 | 1 | 0.36 | -5.38 | -0.29 | 4.74 | 0.36 | -2.89 | 1.34 | 4.61 |
| 4 | *b* igual | 63.4* | 3 | 0.45 | -3.32 | 0.25 | 3.69 | 0.45 | -3.32 | 0.25 | 3.69 |
| 5 | *ab* igual | 25.3* | 4 | 0.52 | -3.81 | -0.65 | 3.21 | 0.35 | -4.33 | -0.91 | 3.62 |
| 5 | *a* igual | 2.9 | 1 | 0.44 | -4.45 | -0.75 | 3.75 | 0.44 | -3.64 | -0.88 | 2.76 |
| 5 | *b* igual | 22.4* | 3 | 0.43 | -4.13 | -0.80 | 3.45 | 0.43 | -4.13 | -0.80 | 3.45 |
| 6 | *ab* igual | 9.9 | 4 | 0.43 | 1.04 | 3.40 | 4.67 | 0.47 | 0.26 | 2.42 | 3.39 |
| 7 | *ab* igual | 5.6 | 4 | 1.07 | -2.01 | -0.58 | 1.62 | 0.93 | -1.92 | -0.50 | 2.02 |
| 8 | *ab* igual | 55.1* | 4 | 0.69 | -2.31 | -0.29 | 1.71 | 0.23 | -2.80 | 1.69 | 7.45 |
| 8 | *a* igual | 20.1* | 1 | 0.47 | -3.22 | -0.39 | 2.41 | 0.47 | -1.77 | 0.49 | 3.36 |
| 8 | *b* igual | 35.0* | 3 | 0.56 | -2.21 | -0.07 | 2.31 | 0.56 | -2.21 | -0.07 | 2.31 |
| 9 | *ab* igual | 10.7 | 4 | 0.81 | -0.98 | 0.73 | 1.95 | 0.62 | -1.60 | 0.84 | 2.58 |
| 10 | *ab* igual | 81.3* | 4 | 1.00 | -1.18 | 0.66 | 3.25 | 0.99 | -1.85 | -0.37 | 2.12 |
| 10 | *a* igual | 0.0 | 1 | 1.00 | -1.19 | 0.66 | 3.27 | 1.00 | -1.84 | -0.38 | 2.10 |
| 10 | *b* igual | 81.3* | 3 | 0.80 | -1.67 | 0.35 | 3.38 | 0.80 | -1.67 | 0.35 | 3.38 |
| 11 | *ab* igual | 65.1* | 4 | 0.95 | -0.41 | 1.23 | 3.69 | 0.84 | 0.59 | 2.05 | 3.59 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 11 | *a* igual | 0.8 | 1 | 0.91 | -0.42 | 1.27 | 3.84 | 0.91 | 0.52 | 1.88 | 3.32 |
| 11 | *b* igual | 64.2* | 3 | 1.01 | -0.07 | 1.35 | 3.36 | 1.01 | -0.07 | 1.35 | 3.36 |
| 12 | *ab* igual | 10.9 | 4 | 1.63 | -1.16 | 0.85 | 2.21 | 1.32 | -1.22 | 0.87 | 2.15 |
| 13 | *ab* igual | 71.2* | 4 | 0.32 | -6.81 | -1.50 | 4.40 | 0.34 | -4.53 | -1.85 | 2.24 |
| 13 | *a* igual | 0.0 | 1 | 0.33 | -6.65 | -1.47 | 4.30 | 0.33 | -4.63 | -1.88 | 2.32 |
| 13 | *b* igual | 71.2 | 3 | 0.30 | -6.10 | -1.74 | 3.83 | 0.30 | -6.10 | -1.74 | 3.83 |
| 14 | *ab* igual | 31.1* | 4 | 0.92 | -0.28 | 1.54 | 3.24 | 1.07 | -0.08 | 1.40 | 2.05 |
| 14 | *a* igual | 1.4 | 1 | 0.98 | -0.26 | 1.46 | 3.07 | 0.98 | -0.04 | 1.54 | 2.25 |
| 14 | *b* igual | 29.8* | 3 | 1.00 | -0.17 | 1.48 | 2.72 | 1.00 | -0.17 | 1.48 | 2.72 |
| 15 | *ab* igual | 6.1 | 4 | 1.30 | -0.81 | 0.82 | 2.08 | 1.01 | -0.97 | 0.95 | 2.39 |
| 15 | *a* igual | 4.9 | 1 | 1.17 | -0.87 | 0.88 | 2.24 | 1.17 | -0.95 | 0.79 | 2.08 |
| 15 | *ab* igual | 1.2 | 3 | 1.15 | -0.91 | 0.86 | 2.22 | 1.15 | -0.91 | 0.86 | 2.22 |
| 16 | *ab* igual | 22.7* | 4 | 0.86 | -1.54 | 1.07 | 2.45 | 1.16 | -1.14 | 0.51 | 1.39 |
| 16 | *a* igual | 5.7 | 1 | 0.98 | -1.39 | 0.96 | 2.20 | 0.98 | -1.19 | 0.67 | 1.67 |
| 16 | *b* igual | 17.1* | 3 | 0.98 | -1.29 | 0.87 | 2.01 | 0.98 | -1.29 | 0.87 | 2.01 |
| 17 | *ab* igual | 32.4* | 4 | 0.25 | -3.01 | -0.13 | 4.56 | 0.36 | -0.83 | 1.41 | 4.60 |
| 17 | *a* igual | 1.3 | 1 | 0.30 | -2.53 | -0.11 | 3.84 | 0.30 | -0.85 | 1.86 | 5.72 |
| 17 | *b* igual | 31.2* | 3 | 0.39 | -1.46 | 0.45 | 3.40 | 0.39 | -1.46 | 0.45 | 3.40 |
| 18 | *ab* igual | 33.3* | 4 | 1.77 | -0.83 | 1.15 | 2.49 | 1.52 | -0.86 | 0.98 | 1.81 |
| 18 | *a* igual | 2.3 | 1 | 1.65 | -0.86 | 1.19 | 2.59 | 1.65 | -0.85 | 0.90 | 1.68 |
| 18 | *b* igual | 31.0* | 3 | 1.59 | -0.86 | 1.13 | 2.31 | 1.59 | -0.86 | 1.13 | 2.31 |
| 19 | *ab* igual | 25.7* | 4 | 1.32 | -1.50 | -0.01 | 2.05 | 1.32 | -1.75 | -0.41 | 1.50 |
| 19 | *a* igual | 0.0 | 1 | 1.32 | -1.50 | -0.01 | 2.06 | 1.32 | -1.75 | -0.41 | 1.50 |
| 19 | *b* igual | 25.7* | 3 | 1.21 | -1.69 | -0.16 | 1.98 | 1.21 | -1.69 | -0.16 | 1.98 |
| 20 | *ab* igual | 13.3* | 4 | 1.17 | -0.75 | 1.64 | 3.04 | 0.97 | -0.57 | 1.75 | 2.95 |
| 20 | *a* igual | 2.2 | 1 | 1.08 | -0.80 | 1.73 | 3.22 | 1.08 | -0.58 | 1.56 | 2.67 |
| 20 | *b* igual | 11.1 | 3 | 1.10 | -0.69 | 1.66 | 2.99 | 1.10 | -0.69 | 1.66 | 2.99 |
| 21 | *ab* igual | 27.2* | 4 | 0.66 | -1.14 | 1.37 | 3.63 | 0.64 | -0.35 | 2.00 | 3.60 |
| 21 | *a* igual | 0.0 | 1 | 0.65 | -1.15 | 1.39 | 3.68 | 0.65 | -0.35 | 1.95 | 3.53 |
| 21 | *b* igual | 27.2* | 3 | 0.72 | -0.76 | 1.43 | 3.28 | 0.72 | -0.76 | 1.43 | 3.28 |
| 22 | *ab* igual | 46.4* | 4 | 1.07 | -0.69 | 0.99 | 3.63 | 1.10 | -0.32 | 0.71 | 2.46 |
| 22 | *a* igual | 0.0 | 1 | 1.08 | -0.68 | 0.98 | 3.60 | 1.08 | -0.32 | 0.72 | 2.49 |
| 22 | *b* igual | 46.3* | 3 | 1.09 | -0.51 | 0.90 | 3.14 | 1.09 | -0.51 | 0.90 | 3.14 |
| 24 | *ab* igual | 32.2* | 4 | 0.93 | -0.50 | 1.76 | 3.98 | 0.80 | -0.15 | 1.81 | 3.24 |
| 24 | *a* igual | 1.3 | 1 | 0.87 | -0.53 | 1.85 | 4.20 | 0.87 | -0.19 | 1.62 | 2.94 |
| 24 | *b* igual | 30.9* | 3 | 0.89 | -0.38 | 1.75 | 3.62 | 0.89 | -0.38 | 1.75 | 3.62 |
| 25 | *ab* igual | 12.9 | 4 | 0.59 | 1.47 | 3.83 | 6.01 | 0.58 | 1.60 | 3.54 | 4.67 |
| 25 | *a* igual | 0.0 | 1 | 0.59 | 1.47 | 3.85 | 6.03 | 0.59 | 1.58 | 3.51 | 4.64 |
| 25 | *b* igual | 12.9* | 3 | 0.59 | 1.52 | 3.73 | 5.44 | 0.59 | 1.52 | 3.73 | 5.44 |
| 26 | *ab* igual | 10.0 | 4 | 0.31 | -1.16 | 3.96 | 6.84 | 0.22 | -1.10 | 7.19 | 12.07 |
| 27 | *ab* igual | 8.7 | 4 | 0.70 | -1.46 | 1.30 | 3.22 | 0.58 | -1.71 | 1.41 | 3.10 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 29 | *ab* igual | 14.9* | 4 | 0.48 | -0.80 | 2.63 | 5.22 | 0.61 | -1.32 | 1.51 | 3.32 |
| 29 | *a* igual | 1.3 | 1 | 0.54 | -0.72 | 2.37 | 4.71 | 0.54 | -1.38 | 1.77 | 3.78 |
| 29 | *b* igual | 13.6* | 3 | 0.47 | -1.09 | 2.45 | 4.95 | 0.47 | -1.09 | 2.45 | 4.95 |

Nota: *$p < 0.01$

Según los resultados de la Tabla 3.52, hay 4 ítems con DIF significativo en referencia al parámetro *a*: el ítem 1, el ítem 2, el ítem 3 y el ítem 8. El ítem que presenta un mayor desajuste en cuanto al parámetro de discriminación es el ítem 2, por lo que se eliminan sus restricciones de igualdad de parámetros entre ambos grupos de edad antes de comparar nuevamente los modelos.

Las diferencias entre ambos modelos siguen siendo significativas ($\Delta G^2$ [26] = 116.3, p <.01) por lo que se eliminan, además, las restricciones de igualdad de parámetros del ítem 8. Sigue sin haber equivalencia entre el modelo sin restricciones y el modelo restringido ($\Delta G^2$ [25] = 95.5, p <.01), por lo que se eliminan las restricciones del ítem 3. De nuevo hay falta de equivalencia entre ambos modelos ($\Delta G^2$ [24] = 87.8, p < .01), por lo que se eliminan las restricciones del ítem 1. No hay equivalencia ($\Delta G^2$ [23] = 77, p < .01), a pesar de haber liberado de la restricción de igualdad de parámetros a todos los ítems que presentaban DIF no uniforme significativo.

Hay dos ítems que, sin llegar a la significación estadística, presentan valores altos de incremento en $G^2$ respecto al parámetro *a*: los ítems 15 y 16. Se elimina la restricción de igualdad del ítem 16 antes de comparar los modelos, encontrando una clara falta de equivalencia ($\Delta G^2$ [22] = 65.9, p < .01). Posteriormente, se hace lo propio con el ítem 15 manteniéndose la falta de equivalencia entre preadolescentes y adolescentes ($\Delta G^2$ [21] = 63.7, p < .01). Además, la diferencia entre los dos últimos modelos no es significativa, por

lo que se detiene el proceso de valoración de modelos anidados, concluyendo que no hay equivalencia entre ambos grupos de edad.

Tabla 3.53. *Equivalencia de medida entre preadolescentes y adolescentes en la escala completa BIS en relación al parámetro de discriminación*

| | $G^2$ | $\Delta G^2$ | $\Delta$ g.l. | $\chi^2$ | $p$ | Ítems libres |
|---|---|---|---|---|---|---|
| Modelo base | 76694.5 | | | | | Todos |
| Modelo invarianza total *a* | 76828.1 | 133.6 | 27 | 47 | .01 | Ninguno |
| Modelo invarianza parcial *a* | 76810.8 | 116.3 | 26 | 45.6 | .01 | Ítem 2 |
| Modelo invarianza parcial *a* | 76790.8 | 95.5 | 25 | 44.3 | .01 | Ítem 2 y 8 |
| Modelo invarianza parcial *a* | 76782.3 | 87.8 | 24 | 43 | .01 | Ítem 2, 8 y 3 |
| Modelo invarianza parcial *a* | 76771.5 | 77 | 23 | 41.6 | .01 | Ítem 2, 8, 3 y 1 |
| Modelo invarianza parcial *a* | 76760.4 | 65.9 | 22 | 40.3 | .01 | Ítem 2, 8, 3, 1 y 16 |
| Modelo invarianza parcial *a* | 76758.2 | 63.7 | 21 | 38.9 | .01 | Ítem 2, 8, 3, 1, 16 y 15 |

Aunque es poco probable encontrar equivalencia en la condición más restrictiva de igualdad de todos los parámetros entre ambos grupos, al no haber equivalencia en la condición de igualdad del parámetro discriminación, se comprueba este hecho a continuación, forzando la igualdad de todos los parámetros de todos los ítems de la subescala ($a$, $b_1$, $b_2$ y $b_3$) y comparando este modelo restringido con el modelo base.

No hay equivalencia total de medida ($\Delta G^2$ [108] = 835.7, $p < .01$). Para establecer si hay equivalencia parcial se utiliza la Tabla 3.52, para eliminar las restricciones del ítem con mayor incremento en $G^2$ antes de volver a comprar el modelo sin restricciones con el modelo base. De esta manera, se eliminan sucesivamente las restricciones de los ítems 10, 13, 11, 4, 8, 3, 22, 2, 1, 18, 17, 24, 14, 21, 19, 5, 29 y 20 (ver Tabla 3.54 para consultar los valores de los índices calculados), y aún eliminando la restricción de igualdad entre los

parámetros de los 18 ítems que presentan DIF en la escala sigue habiendo diferencias significativas entre el modelo base y el modelo que restringe la igualdad de los parámetros de 9 ítems entre ambos grupos de edad. Además la mejora del ajuste producida entre los dos últimos modelos anidados, que restringen la igualdad de parámetros de 18 y 17 ítems respectivamente ya no es significativa, con lo que no tiene sentido continuar con el proceso. La escala BIS presenta funcionamiento diferencial en relación a la edad.

Tabla 3.54. *Equivalencia de medida entre preadolescentes y adolescentes en la escala BIS forzando la igualdad de todos los parámetros*

| | $G^2$ | $\Delta G^2$ | $\Delta$ g.l. | $\chi^2$ | $p$ | Ítems libres |
|---|---|---|---|---|---|---|
| Modelo base | 76694.5 | | | | | Todos |
| Modelo invarianza total *a* y *b* | 77530.2 | 835.7 | 108 | 135.81 | .01 | Ninguno |
| Modelo invarianza parcial *a* y *b* | 77448.1 | 753.6 | 104 | 135.81 | .01 | Ítem 10 |
| Modelo invarianza parcial *a* y *b* | 77376.2 | 681.7 | 100 | 135.81 | .01 | y 13 |
| Modelo invarianza parcial *a* y *b* | 77316.1 | 621.6 | 96 | 135.81 | .01 | y 11 |
| Modelo invarianza parcial *a* y *b* | 77251.7 | 557.2 | 92 | 124.12 | .01 | y 4 |
| Modelo invarianza parcial *a* y *b* | 77197.2 | 502.7 | 88 | 124.12 | .01 | y 8 |
| Modelo invarianza parcial *a* y *b* | 77153.5 | 459 | 84 | 112.33 | .01 | y 3 |
| Modelo invarianza parcial *a* y *b* | 77107.4 | 412.9 | 80 | 112.33 | .01 | y 22 |
| Modelo invarianza parcial *a* y *b* | 77064.7 | 370.2 | 76 | 112.33 | .01 | y 2 |
| Modelo invarianza parcial *a* y *b* | 77029.4 | 334.9 | 72 | 100.42 | .01 | y 1 |
| Modelo invarianza parcial *a* y *b* | 76998.4 | 303.9 | 68 | 100.42 | .01 | y 18 |
| Modelo invarianza parcial *a* y *b* | 76964.8 | 270.3 | 64 | 88.38 | .01 | y 17 |
| Modelo invarianza parcial *a* y *b* | 76931.9 | 237.4 | 60 | 88.38 | .01 | y 24 |
| Modelo invarianza parcial *a* y *b* | 76897.8 | 203.3 | 56 | 88.38 | .01 | y 14 |
| Modelo invarianza parcial *a* y *b* | 76867.7 | 173.2 | 52 | 76.15 | .01 | y 21 |
| Modelo invarianza parcial *a* y *b* | 76848.0 | 153.5 | 48 | 76.15 | .01 | y 19 |
| Modelo invarianza parcial *a* y *b* | 76823.0 | 128.5 | 44 | 63.69 | .01 | y 5 |
| Modelo invarianza parcial *a* y *b* | 76809.5 | 115 | 40 | 63.69 | .01 | y 29 |
| Modelo invarianza parcial *a* y *b* | 76796.0 | 101.5 | 36 | 63.69 | .01 | y 20 |

Nota: a partir del primer modelo de invarianza parcial, los ítems libres serán el indicado en la celdilla correspondiente más los reflejados en las filas anteriores de la misma columna.

Los resultados sobre qué ítems presentan DIF en el test BIS son muy similares a los encontrados en el análisis relativo a cada subescala por separado:

(1) En la subescala Impulso Motor en ambos casos se detectan 3 ítems, pero hay uno que no es coincidente: el ítem 29 presenta DIF en el análisis de la escala completa,

pero no en el de las subescalas, y el ítem 12 es el caso contrario, ya que presenta DIF en el análisis de la subescala, pero no en de la escala completa.

(2) En la subescala Impulso No Planificado a nivel del test completo se detecta un ítem más (el ítem 25) que en el análisis de la escala completa, coincidiendo los 8 ítems restantes con DIF.

(3) En la subescala Impulso Cognitivo-Atencional coinciden los mismos 7 ítems con DIF, detectándose en el caso del análisis del test completo un ítem más: el número 19.

A pesar del elevado número de ítems con DIF, detectado por el procedimiento de comparación de modelos basado en el test LR, la CCT de ambos grupos de edad no refleja este hecho, debido posiblemente a que el DIF de los distintos ítems tienen direcciones opuestas, y a nivel de test estos efectos pueden verse cancelados (ver Figura 3.30).

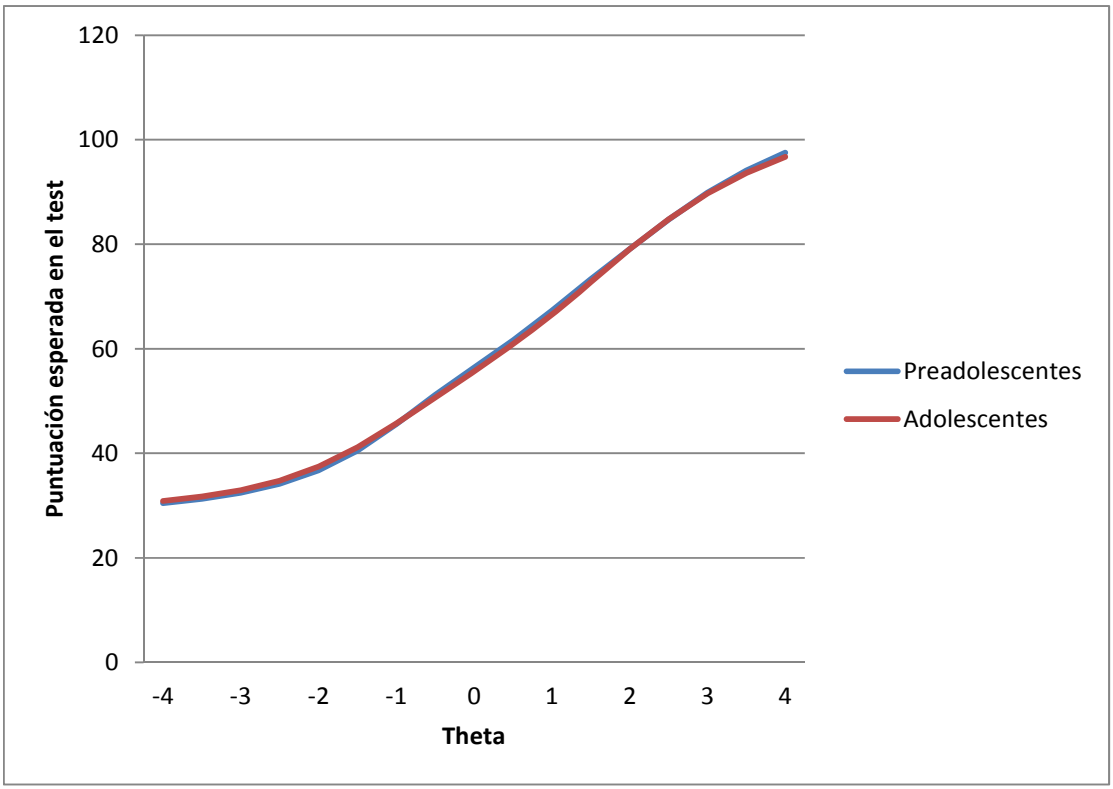



Figura 3.30. CCT para ambos grupos de edad en el test BIS.

## 3.5. INVARIANZA MEDIANTE EL PROCEDIMIENTO DFIT

### 3.5.1. EQUIVALENCIA DE MEDIDA ENTRE HOMBRES Y MUJERES

#### *3.5.1.1. Subescala Impulso Motor del BIS*

En primer lugar, se estiman los parámetros de los ítems de la subescala de Impulso Motor de BIS-PA para hombres y mujeres por separado (ver tablas 3.55 y 3.56)

Tabla 3.55. *Parámetros estimados (y errores estándar asociados) para los ítems de la subescala Impulso Motor del BIS en la muestra de hombres*

| Item | *a* | *b1* | *b2* | *b3* |
|---|---|---|---|---|
| 2 | 1.69 (0.10) | -0.83 (0.07) | 1.31 (0.09) | 2.76 (0.19) |
| 6 | 0.37 (0.08) | 1.39 (0.38) | 3.73 (0.88) | 5.15 (1.22) |
| 9 | 0.71 (0.08) | -1.09 (0.17) | 0.99 (0.16) | 2.43 (0.28) |
| 12 | 1.88 (0.11) | -0.73 (0.07) | 1.01 (0.07) | 2.06 (0.12) |
| 15 | 1.17 (0.09) | -0.61 (0.09) | 1.22 (0.11) | 2.51 (0.20) |
| 18 | 1.94 (0.11) | -0.46 (0.06) | 1.16 (0.08) | 2.25 (0.14) |
| 26 | 0.28 (0.07) | -0.82 (0.32) | 5.24 (1.34) | 9.14 (2.47) |
| 29 | 0.48 (0.08) | -0.29 (0.19) | 2.95 (0.48) | 5.31 (0.85) |

Nota: los errores estándar aparecen entre paréntesis.

Tabla 3.56. *Parámetros estimados (y errores estándar asociados) para los ítems de la subescala Impulso Motor del BIS en la muestra de mujeres*

| Item | *a* | *b1* | *b2* | *b3* |
|---|---|---|---|---|
| 2 | 1.89 (0.11) | -0.65 (0.06) | 1.31 (0.08) | 2.71 (0.17) |
| 6 | 0.40 (0.07) | 1.00 (0.26) | 3.62 (0.71) | 4.91 (0.96) |
| 9 | 0.73 (0.08) | -0.69 (0.13) | 1.23 (0.16) | 2.50 (0.28) |
| 12 | 2.46 (0.12) | -0.57 (0.05) | 1.02 (0.06) | 2.00 (0.10) |
| 15 | 1.41 (0.10) | -0.34 (0.07) | 1.09 (0.08) | 2.22 (0.15) |
| 18 | 2.23 (0.13) | -0.33 (0.05) | 1.32 (0.07) | 2.16 (0.12) |
| 26 | 0.54 (0.09) | -0.46 (0.11) | 2.75 (0.07) | 4.30 (0.12) |
| 29 | 0.66 (0.08) | -0.77(0.15) | 2.09 (0.25) | 3.70 (0.44) |

Nota: los errores estándar aparecen entre paréntesis.

Antes de analizar la equivalencia de medida entre hombres y mujeres en Impulso Motor hay que igualar los parámetros de los ítems del grupo focal a la métrica subyacente de los parámetros de los ítems del grupo de referencia. Se considera como grupo de referencia el grupo de mujeres porque cuenta con un mayor número de participantes.

Los coeficientes de transformación métrica calculados con el programa EQUATE son:

A = 0.8397 y K = 0.0794.

Dado que:

$$a_i^* = \frac{a_i}{A}$$

y

$$b_{i1}^* = Ab_{i1} + K$$

$$\vdots$$

$$b_{im-1}^* = Ab_{im-1} + K$$

donde:

*A* es la pendiente

*K* es la pendiente en el origen

$a_i$ es el índice de discriminación del ítem *i* en el grupo focal antes de transformarse en la métrica del grupo de referencia

$a^*_i$ es el parámetro de discriminación del ítem *i* en el grupo focal expresado en la misma métrica que el grupo de referencia

$b_{i1}$ es el parámetro de umbral de la categoría 1 a la categoría 2 del ítem *i* en el grupo focal antes de transformarse en la métrica del grupo de referencia

$b_{im-1}$ es el parámetro de umbral de la categoría *m-1* a la categoría *m* del ítem *i* en el grupo focal antes de transformarse en la métrica del grupo de referencia

$b^*_{im-1}$ es el parámetro de umbral de la categoría *m-1* a la categoría *m* del ítem *i* en el grupo expresado en la misma métrica que el grupo de referencia.

Con estos coeficientes se transforman los parámetros del grupo de hombres en la métrica subyacente del grupo de mujeres. Una vez hecho esto, ya se puede comparar la equivalencia de medida entre ambos grupos con DFIT8.

En la tabla 3.57 se muestran los parámetros de hombres y mujeres en la subescala Impulso Motor del BIS-PA en la misma métrica, acompañados de una medida del

funcionamiento diferencial compensatorio del ítem (CDIF), y una medida del funcionamiento diferencial no compensatorio del ítem (NCDIF), junto al punto de corte para este índice, establecido con un nivel de significación de .01.

Tabla 3.57. *Parámetros estimados del ítem para hombres y mujeres en la subescala Impulso Motor, e índices de funcionamiento diferencial del ítem*

| Item | *a* | *b1* | *b2* | *b3* | *CDIF* | *NCDIF* | Pto. corte* | Sig. |
|---|---|---|---|---|---|---|---|---|
| Item 2 | | | | | 0.00088 | 0.00049 | 0.00266 | ns |
| Hombres | 2.01 | -0.62 | 1.18 | 2.39 | | | | |
| Mujeres | 1.89 | -0.65 | 1.31 | 2.71 | | | | |
| Item 6 | | | | | 0.00118 | 0.00110 | 0.01027 | ns |
| Hombres | 0.44 | 1.25 | 3.21 | 4.40 | | | | |
| Mujeres | 0.41 | 1.00 | 3.62 | 4.91 | | | | |
| Item 9 | | | | | -0.00049 | 0.00240 | 0.00767 | ns |
| Hombres | 0.85 | -0.84 | 0.91 | 2.12 | | | | |
| Mujeres | 0.73 | -0.69 | 1.23 | 2.50 | | | | |
| Item 12 | | | | | 0.00035 | 0.00051 | 0.00204 | ns |
| Hombres | 2.23 | -0.54 | 0.92 | 1.81 | | | | |
| Mujeres | 2.46 | -0.57 | 1.02 | 2.00 | | | | |
| Item 15 | | | | | -0.00114 | 0.00098 | 0.00410 | ns |
| Hombres | 1.39 | -0.43 | 1.10 | 2.19 | | | | |
| Mujeres | 1.41 | -0.34 | 1.09 | 2.22 | | | | |
| Item 18 | | | | | 0.00108 | 0.00101 | 0.00248 | ns |
| Hombres | 2.31 | -0.30 | 1.06 | 1.97 | | | | |
| Mujeres | 2.23 | -0.33 | 1.32 | 2.16 | | | | |
| Item 26 | | | | | -0.00166 | 0.00363 | 0.00720 | ns |
| Hombres | 0.34 | -0.61 | 4.48 | 7.76 | | | | |
| Mujeres | 0.54 | -0.46 | 2.75 | 4.30 | | | | |
| Item 29 | | | | | 0.00259 | 0.00968 | 0.00632 | .001 |
| Hombres | 0.57 | -0.17 | 2.55 | 4.54 | | | | |
| Mujeres | 0.66 | -0.77 | 2.09 | 3.70 | | | | |

*Punto de corte establecido del índice NCDIF con el procedimiento IPR para α =.01

Considerando los ítems de la subescala individualmente (ver tabla 3.57) únicamente el ítem 29 presenta funcionamiento diferencial, ya que su valor de NCDIF es superior al

punto de corte establecido para ese ítem por el procedimiento IPR, teniendo en cuenta un α de .01.

La suma de los índices de funcionamiento diferencial compensatorio (CDIF) de todos los ítems de la subescala proporciona el valor del funcionamiento diferencial de la subescala completa, siendo en este caso 0.00279. Dado que este valor es muy inferior al punto de corte establecido por el procedimiento (0.04568) no hay funcionamiento diferencial de la subescala, por lo que no es necesario eliminar ninguno de sus ítems, y queda probada mediante este procedimiento la equivalencia entre hombres y mujeres en la subescala de Impulso Motor.

En la figura 3.31 se muestra la curva característica del test para ambos sexos, observándose que ambas líneas están prácticamente superpuestas en todos los niveles de $\theta$, exceptuando el nivel intermedio-alto de impulso motor, con una puntuación esperada ligeramente mayor para los hombres que para las mujeres..

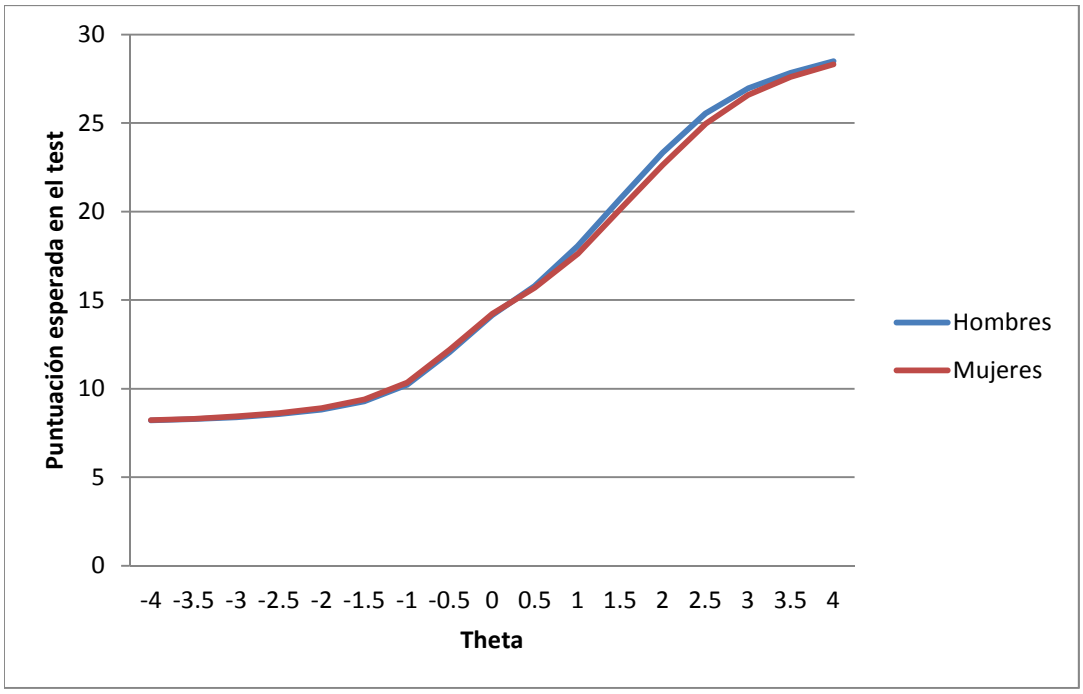



Figura 3.31. CCT en hombres y mujeres en la subescala de Impulso Motor.

### *3.5.1.2. Subescala Impulso no Planificado del BIS*

Las estimaciones de los parámetros de los ítems de la escala Impulso No Planificado en hombres y en mujeres se muestran en las tablas 3.58 y 3.59.

Tabla 3.58. *Parámetros estimados (y errores estándar asociados) para los ítems de la subescala Impulso No Planificado del BIS en la muestra de hombres*

| Item | *a* | *b1* | *b2* | *b3* |
|---|---|---|---|---|
| 1 | 0.99 (0.09) | -1.74 (0.17) | -0.01 (0.09) | 3.64 (0.36) |
| 3 | 0.85 (0.08) | -1.31 (0.16) | 1.58 (0.18) | 3.57 (0.37) |
| 5 | 0.50 (0.07) | -3.21 (0.49) | -0.48 (0.18) | 2.72 (0.44) |
| 8 | 0.64 (0.08) | -1.55 (0.23) | 0.44 (0.14) | 2.33 (0.31) |
| 11 | 1.41 (0.11) | 0.02 (0.07) | 1.15 (0.09) | 2.61 (0.20) |
| 14 | 1.36 (0.10) | 0.11 (0.07) | 1.22 (0.10) | 2.23 (0.17) |
| 17 | 0.61 (0.07) | -0.75 (0.17) | 0.41 (0.15) | 2.02 (0.29) |
| 22 | 1.38 (0.10) | -0.20 (0.07) | 1.05 (0.09) | 2.66 (0.21) |
| 25 | 0.40 (0.09) | 1.93 (0.46) | 4.88 (1.11) | 6.99 (1.60) |

Nota: los errores estándar aparecen entre paréntesis.

Tabla 3.59. *Parámetros estimados (y errores estándar asociados) para los ítems de la subescala Impulso No Planificado del BIS en la muestra de mujeres*

| Item | *a* | *b1* | *b2* | *b3* |
|---|---|---|---|---|
| 1 | 1.03 (0.09) | -1.25 (0.13) | 0.14 (0.09) | 3.99 (0.38) |
| 3 | 1.11 (0.10) | -0.55 (0.09) | 1.91 (0.16) | 3.48 (0.32) |
| 5 | 0.39 (0.08) | -3.95 (0.75) | -0.62 (0.25) | 4.09 (0.83) |
| 8 | 0.66 (0.08) | -1.74 (0.24) | -0.04 (0.13) | 2.06 (0.27) |
| 11 | 1.24 (0.11) | 0.31 (0.08) | 1.51 (0.13) | 3.10 (0.28) |
| 14 | 1.40 (0.10) | 0.29 (0.07) | 1.61 (0.12) | 2.49 (0.19) |
| 17 | 0.49 (0.08) | -0.69 (0.21) | 1.06 (0.24) | 3.72 (0.61) |
| 22 | 1.97 (0.12) | -0.02(0.05) | 0.81 (0.06) | 2.17 (0.13) |
| 25 | 0.69 (0.10) | 1.84 (0.27) | 4.16 (0.60) | 5.47 (0.84) |

Nota: los errores estándar aparecen entre paréntesis.

Puesto que los parámetros de los ítems estimados en las tablas 3.58 y 3.59 tienen una métrica subyacente distinta es necesario igualar la métrica del grupo de hombres (grupo focal) a la métrica del grupo de mujeres (grupo de referencia). Los coeficientes de transformación son: A = 0.9856 y K = 0.2544

Una vez transformados los parámetros del grupo de hombres se analiza la equivalencia de medida entre ambos grupos. En la tabla 3.60 se muestran los parámetros de cada uno de los ítems de la subescala para ambos grupos, así como los índices de equivalencia CDIF, NCDIF y el punto de corte asociado a este último.

Tabla 3.60. *Parámetros estimados del ítem para hombres y mujeres en la subescala Impulso No Planificado, e índices de funcionamiento diferencial del ítem*

| Item | *a* | *b1* | *b2* | *b3* | *CDIF* | *NCDIF* | Pto. corte* | Sig. |
|---|---|---|---|---|---|---|---|---|
| Item 1 | | | | | 0.00267 | 0.00098 | 0.00598 | ns |
| Hombres | 1.00 | -1.46 | 0.25 | 3.84 | | | | |
| Mujeres | 1.03 | -1.25 | 0.14 | 3.99 | | | | |
| Item 3 | | | | | 0.01630 | 0.02812 | 0.00348 | .001 |
| Hombres | 0.86 | -1.03 | 1.82 | 3.78 | | | | |
| Mujeres | 1.11 | -0.55 | 1.91 | 3.48 | | | | |
| Item 5 | | | | | -0.00711 | 0.00683 | 0.00999 | ns |
| Hombres | 0.51 | -2.91 | -0.22 | 2.94 | | | | |
| Mujeres | 0.40 | -3.95 | -0.62 | 4.09 | | | | |
| Item 8 | | | | | -0.01971 | 0.04217 | 0.00968 | .001 |
| Hombres | 0.65 | -1.27 | 0.69 | 2.56 | | | | |
| Mujeres | 0.66 | -1.74 | -0.04 | 2.06 | | | | |
| Item 11 | | | | | -0.00309 | 0.00102 | 0.00332 | ns |
| Hombres | 1.43 | 0.27 | 1.39 | 2.83 | | | | |
| Mujeres | 1.24 | 0.31 | 1.51 | 3.10 | | | | |
| Item 14 | | | | | 0.00007 | 0.00003 | 0.00273 | ns |
| Hombres | 1.39 | 0.36 | 1.45 | 2.45 | | | | |
| Mujeres | 1.40 | 0.29 | 1.61 | 2.49 | | | | |
| Item 17 | | | | | 0.00107 | 0.00290 | 0.00991 | ns |
| Hombres | 0.62 | -0.48 | 0.66 | 2.24 | | | | |
| Mujeres | 0.49 | -0.69 | 1.06 | 3.72 | | | | |
| Item 22 | | | | | -0.00377 | 0.00544 | 0.00338 | .001 |
| Hombres | 1.40 | 0.06 | 1.29 | 2.88 | | | | |
| Mujeres | 1.97 | -0.03 | 0.81 | 2.17 | | | | |
| Item 25 | | | | | 0.01627 | 0.02809 | 0.00254 | .001 |
| Hombres | 0.41 | 2.16 | 5.06 | 7.14 | | | | |
| Mujeres | 0.69 | 1.84 | 4.16 | 5.47 | | | | |

*Punto de corte establecido del índice NCDIF con el procedimiento IPR para α =.01

Cuatro de los ítems de la subescala INP presentan funcionamiento diferencial del ítem, porque el valor de NCDIF es mayor que el punto de corte referenciado: los ítems 3, 8, 22 y 25 (ver Tabla 3.60). Según el índice CDIF, dos de estos ítems presentan DIF a favor del grupo focal y dos a favor del grupo de referencia, por lo que cabe esperar que a nivel del test, se compensen en buena medida.

En efecto, el índice DTF arrojó un valor de 0.01024, muy lejos del punto de corte establecido en 0.04568, por lo que se puede concluir que existe equivalencia de medida entre hombres y mujeres en Impulso No Planificado, sin resultar necesario eliminar ningún ítem de la subescala para que esto suceda.

Se puede apreciar como la como la curva característica del test para ambos sexos, mostrada en la Figura 3.32 es consistente con el resultado numérico, observándose que ambas líneas están prácticamente superpuestas en todos los niveles de $\theta$.

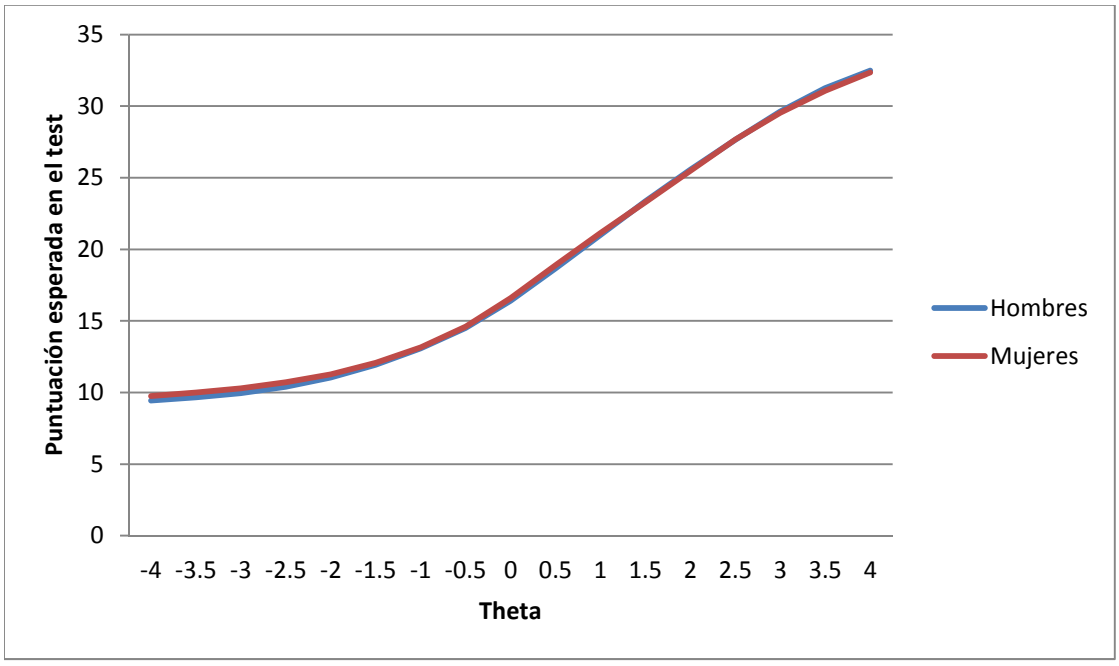



Figura 3.32. CCT en hombres y mujeres en la subescala Impulso No Planificado

### *3.5.1.3. Subescala Impulso Cognitivo-Atencional del BIS*

Las estimaciones de los parámetros para hombres y mujeres en la escala de Impulso Cognitivo-Atencional se muestran en las Tablas 3.61 y 3.62.

Tabla 3.61. *Parámetros estimados (y errores estándar asociados) para los ítems de la subescala Impulso Cognitivo-Atencional del BIS en la muestra de hombres*

| Item | *a* | *b1* | *b2* | *b3* |
|---|---|---|---|---|
| 4 | 0.20 (0.07) | -5.74 (2.14) | 1.07 (0.57) | 8.45 (3.13) |
| 7 | 1.23 (0.08) | -1.46 (0.12) | 0.22 (0.08) | 1.89 (0.14) |
| 10 | 0.83 (0.08) | -1.47 (0.18) | -0.48 (0.11) | 3.28 (0.35) |
| 13 | 0.54 (0.08) | -2.81 (0.42) | -0.50 (0.17) | 2.46 (0.35) |
| 16 | 1.19 (0.09) | -0.69 (0.09) | 0.95 (0.10) | 1.83 (0.15) |
| 19 | 1.23 (0.10) | -1.43 (0.12) | 0.08 (0.07) | 2.03 (0.15) |
| 20 | 1.18 (0.10) | -0.34 (0.08) | 1.70 (0.14) | 2.91 (0.24) |
| 21 | 0.56 (0.08) | -0.30(0.16) | 2.09 (0.32) | 4.37 (0.64) |
| 24 | 0.92 (0.09) | -0.16 (0.10) | 1.68 (0.18) | 3.18 (0.33) |
| 27 | 0.38 (0.07) | -2.05 (0.46) | 2.08 (0.44) | 4.94 (0.99) |

Nota: los errores estándar aparecen entre paréntesis.

Tabla 3.62. *Parámetros estimados (y errores estándar asociados) para los ítems de la subescala Impulso Cognitivo-Atencional del BIS en la muestra de mujeres*

| Item | *a* | *b1* | *b2* | *b3* |
|---|---|---|---|---|
| 4 | 0.37 (0.07) | -3.80 (0.82) | -0.76 (0.28) | 4.86 (1.03) |
| 7 | 1.55 (0.11) | -1.22 (0.09) | -0.24 (0.06) | 1.54 (0.11) |
| 10 | 0.91 (0.08) | -0.91 (0.13) | 0.77 (0.12) | 3.32 (0.33) |
| 13 | 0.59 (0.07) | -3.30 (0.45) | -0.95 (0.19) | 2.08 (0.30) |
| 16 | 1.54 (0.10) | -0.72 (0.07) | 0.91 (0.08) | 1.76 (0.12) |
| 19 | 1.02 (0.08) | -1.26 (0.13) | 0.30 (0.09) | 2.56 (0.23) |
| 20 | 1.32 (0.10) | -0.15 (0.07) | 1.71 (0.13) | 2.81 (0.22) |
| 21 | 0.77 (0.09) | -0.33 (0.12) | 1.78 (0.21) | 3.29 (0.37) |
| 24 | 1.03 (0.09) | 0.11 (0.09) | 2.01 (0.18) | 3.66 (0.35) |
| 27 | 0.56 (0.08) | -1.08 (0.21) | 2.25 (0.34) | 4.41 (0.63) |

Nota: los errores estándar aparecen entre paréntesis.

Una vez estimados los parámetros del ítem para hombres y mujeres se calculan los coeficientes de transformación para igualar la métrica del grupo de hombres al grupo de mujeres. Estos coeficientes de transformación son: A = 0.8854 y K = 0.1347.

Los parámetros del ítem estimados para hombres y mujeres, una vez igualada la métrica del grupo focal (hombres) a la del grupo de referencia (mujeres), se muestran en la Tabla 3.63, junto con los índices de equivalencia.

Tabla 3.63. *Parámetros estimados del ítem para hombres y mujeres en la subescala Impulso Cognitivo-Atencional, e índices de funcionamiento diferencial del ítem*

| Item | *a* | *b1* | *b2* | *b3* | *CDIF* | *NCDIF* | Pto. Corte* | *Sig.* |
|---|---|---|---|---|---|---|---|---|
| Item 4 | | | | | -0.00032 | 0.00152 | 0.00737 | ns |
| Hombres | 0.25 | -4.95 | 1.08 | 7.62 | | | | |
| Mujeres | 0.37 | -3.80 | 0.76 | 4.86 | | | | |
| Item 7 | | | | | 0.00002 | 0.00874 | 0.00617 | .005 |
| Hombres | 1.39 | -1.16 | -0.06 | 1.81 | | | | |
| Mujeres | 1.55 | -1.22 | -0.24 | 1.54 | | | | |
| Item 10 | | | | | -0.00012 | 0.00888 | 0.00577 | .001 |
| Hombres | 0.94 | -1.16 | 0.56 | 3.04 | | | | |
| Mujeres | 0.91 | -0.91 | 0.77 | 3.32 | | | | |
| Item 13 | | | | | 0.00046 | 0.04038 | 0.00800 | .001 |
| Hombres | 0.61 | -2.36 | -0.31 | 2.31 | | | | |
| Mujeres | 0.59 | -3.30 | -0.95 | 2.08 | | | | |
| Item 16 | | | | | 0.00017 | 0.00271 | 0.00353 | ns |
| Hombres | 1.35 | -0.47 | 0.97 | 1.76 | | | | |
| Mujeres | 1.54 | -0.72 | 0.91 | 1.76 | | | | |
| Item 19 | | | | | 0.00070 | 0.00503 | 0.00593 | ns |
| Hombres | 1.39 | -1.13 | 0.20 | 1.93 | | | | |
| Mujeres | 1.02 | -1.26 | 0.30 | 2.56 | | | | |
| Item 20 | | | | | 0.00007 | 0.00009 | 0.00347 | ns |
| Hombres | 1.34 | -0.17 | 1.64 | 2.71 | | | | |
| Mujeres | 1.32 | -0.16 | 1.71 | 2.81 | | | | |
| Item 21 | | | | | -0.00033 | 0.00081 | 0.00532 | ns |
| Hombres | 0.63 | -0.14 | 1.98 | 4.00 | | | | |
| Mujeres | 0.77 | -0.33 | 1.78 | 3.29 | | | | |
| Item 24 | | | | | 0.00022 | 0.00685 | 0.00324 | .001 |
| Hombres | 1.04 | -0.01 | 1.62 | 2.95 | | | | |
| Mujeres | 1.03 | 0.11 | 2.01 | 3.66 | | | | |
| Item 27 | | | | | -0.00039 | 0.02790 | 0.00584 | .001 |
| Hombres | 0.43 | -1.68 | 1.92 | 4.51 | | | | |
| Mujeres | 0.56 | -1.08 | 2.25 | 4.44 | | | | |

*Punto de corte establecido del índice NCDIF con el procedimiento IPR para α =.01

Habiendo fijado un nivel máximo de significación de .01, son 5 los ítems que presentan DIF, superando su valor de NCDIF el punto de corte establecido: los ítems 7, 10, 13, 24 y 27 (ver Tabla 3.63).

En cuanto al funcionamiento diferencial de la escala completa, el índice DTF presenta un valor de 0.00050, muy lejos del 0.05484 que marcaría el funcionamiento diferencial del test según la estimación proporcionada por el procedimiento IPR. Por tanto, existe equivalencia de medida entre hombres y mujeres a nivel de subescala, en el caso de Impulso Cognitivo Atencional del BIS, aunque 5 de los 10 ítems de la subescala presentan funcionamiento diferencial. De los ítems con DIF, los hombres precisan de un mayor nivel de impulsividad para obtener la misma probabilidad de marcar una misma opción de respuesta en los ítems 7, 13 y 24, precisando mayores niveles de rasgo las mujeres en los ítems 10 y 27.

En la curva característica del test para hombres y mujeres (ver Figura 3.33) ambas líneas están superpuestas en todos los niveles de ICA, apreciándose la equivalencia de medida encontrada.

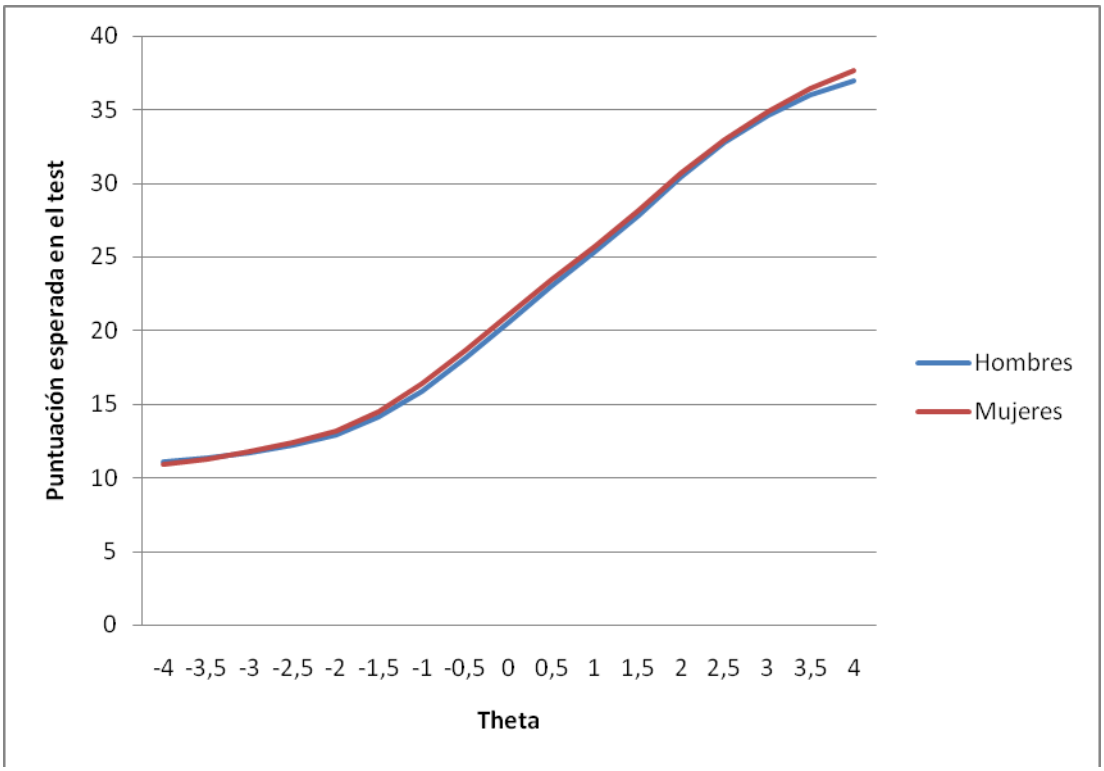


Figura 3.33 CCT en hombres y mujeres en la subescala de Impulso Cognitivo-Atencional.

#### *3.5.1.4. Escala total BIS-PA*

Se estimaron los parámetros de los ítems de la escala para hombres y mujeres por separado (ver tablas 3.64 y 3.65)

Tabla 3.64. *Parámetros estimados (y errores estándar asociados) para los ítems del BIS en la muestra de hombres*

| Item | *a* | *b1* | *b2* | *b3* |
|---|---|---|---|---|
| 1 | 0.98 (0.09) | -1.78 (0.18) | -0.04 (0.10) | 3.64 (0.38) |
| 2 | 1.36 (0.10) | -0.96 (0.09) | 1.44 (0.12) | 3.11 (0.26) |
| 3 | 1.04 (0.09) | -1.15 (0.13) | 1.33 (0.14) | 3.01 (0.28) |
| 4 | 0.38 (0.08) | -3.10 (0.87) | 0.55 (0.25) | 4.49 (0.91) |
| 5 | 0.52 (0.08) | -3.12 (0.48) | -0.50 (0.18) | 2.58 (0.43) |
| 6 | 0.41 (0.09) | 1.25 (0.34) | 3.38 (0.76) | 4.67 (1.05) |
| 7 | 1.03 (0.09) | -1.66 (0.16) | -0.27 (0.09) | 2.10 (0.19) |
| 8 | 0.50 (0.08) | -1.94 (0.35) | -0.52 (0.19) | 2.86 (0.48) |
| 9 | 0.73 (0.09) | -1.10 (0.18) | 0.95 (0.16) | 2.36 (0.29) |
| 10 | 0.80 (0.08) | -1.52 (0.19) | 0.48 (0.12) | 3.36 (0.36) |
| 11 | 0.96 (0.09) | -0.01 (0.10) | 1.44 (0.16) | 3.40 (0.35) |
| 12 | 1.29 (0.10) | -0.92 (0.10) | 1.19 (0.11) | 2.49 (0.20) |
| 13 | 0.33 (0.07) | -4.53 (1.03) | -0.81 (0.32) | 3.92 (0.97) |
| 14 | 1.12 (0.10) | 0.08 (0.08) | 1.33 (0.13) | 2.50 (0.22) |
| 15 | 1.07 (0.09) | -0.67 (0.10) | 1.26 (0.13) | 2.64 (0.24) |
| 16 | 0.99 (0.09) | -0.80 (0.12) | 1.04 (0.12) | 2.05 (0.19) |
| 17 | 0.49 (0.08) | -0.94 (0.24) | 0.46 (0.20) | 2.42 (0.42) |
| 18 | 1.61 (0.11) | -0.51 (0.07) | 1.24 (0.09) | 2.42 (0.18) |
| 19 | 1.25 (0.09) | -1.44 (0.12) | 0.07 (0.08) | 2 (0.16) |
| 20 | 1.15 (0.10) | -0.37 (0.09) | 1.71 (0.15) | 2.94 (0.25) |
| 21 | 0.71 (0.09) | -0.27 (0.13) | 1.67 (0.23) | 3.49 (0.44) |
| 22 | 0.97 (0.09) | -0.28 (0.10) | 1.27 (0.15) | 3.38 (0.36) |
| 24 | 0.94 (0.09) | -0.18 (0.10) | 1.63 (0.17) | 3.09 (0.32) |
| 25 | 0.51 (0.10) | 1.51 (0.31) | 3.85 (0.71) | 5.52 (1.03) |
| 26 | 0.26 (0.25) | -0.93 (0.80) | 5.49 (2.37) | 9.63 (4.31) |
| 27 | 0.61 (0.08) | -1.37 (0.23) | 1.32 (0.22) | 3.23 (0.44) |
| 29 | 0.47 (0.08) | -0.33 (0.20) | 2.93 (0.50) | 5.34 (0.89) |

Nota: los errores estándar aparecen entre paréntesis.

Tabla 3.65. *Parámetros estimados (y errores estándar asociados) para los ítems del BIS en la muestra de mujeres*

| Item | *a* | *b1* | *b2* | *b3* |
|---|---|---|---|---|
| 1 | 0.99 (0.08) | -1.37 (0.13) | 0.15 (0.09) | 4.11 (0.40) |
| 2 | 1.49 (0.10) | -0.79 (0.07) | 1.43 (0.10) | 3.04 (0.22) |
| 3 | 1.19 (0.09) | -0.66 (0.08) | 1.76 (0.14) | 3.23 (0.26) |
| 4 | 0.49 (0.08) | -2.88 (0.47) | 0.57 (0.19) | 3.78 (0.61) |
| 5 | 0.38 (0.07) | -4.31 (0.85) | -0.57 (0.23) | 4.34 (0.86) |
| 6 | 0.46 (0.08) | 0.84 (0.22) | 3.16 (0.57) | 4.29 (0.77) |
| 7 | 1.29 (0.09) | -1.40 (0.11) | -0.28 (0.07) | 1.69 (0.12) |
| 8 | 0.66 (0.07) | -1.74 (0.22) | -0.04 (0.12) | 2 (0.25) |
| 9 | 0.72 (0.08) | -0.75 (0.14) | 1.19 (0.16) | 2.47 (0.28) |
| 10 | 0.89 (0.09) | -1.13 (0.13) | 0.64 (0.10) | 3.30 (0.27) |
| 11 | 1.04 (0.09) | 0.35 (0.08) | 1.68 (0.16) | 3.51 (0.34) |
| 12 | 1.63 (0.10) | -0.72 (0.07) | 1.16 (0.08) | 2.34 (0.15) |
| 13 | 0.39 (0.07) | -4.87 (0.88) | -1.42 (0.31) | 2.92 (0.57) |
| 14 | 1.16 (0.10) | 0.22 (0.07) | 1.80 (0.15) | 2.89 (0.24) |
| 15 | 1.37 (0.10) | -0.39 (0.07) | 1.06 (0.09) | 2.21 (0.16) |
| 16 | 1.16 (0.09) | -0.92 (0.10) | 1.03 (0.10) | 2.09 (0.17) |
| 17 | 0.33 (0.07) | -1.17 (0.35) | 1.49 (0.41 | 5.46 (1.20) |
| 18 | 1.78 (0.11) | -0.41 (0.06) | 1.40 (0.09) | 2.34 (0.15) |
| 19 | 1.21 (0.09) | -1.17 (0.10) | 0.15 (0.07) | 2.22 (0.17) |
| 20 | 1.28 (0.09) | -0.20 (0.07) | 1.77 (0.13) | 2.83 (0.22) |
| 21 | 0.93 (0.09) | -0.40 (0.09) | 1.42 (0.15) | 2.81 (0.26) |
| 22 | 1.21 (0.09) | -0.13 (0.07) | 0.97 (0.10) | 2.93 (0.25) |
| 24 | 0.97 (0.09) | 0.06 (0.09) | 2.07 (0.19) | 3.69 (0.36) |
| 25 | 0.85 (0.11) | 1.57 (0.19) | 3.48 (0.41) | 4.65 (0.61) |
| 26 | 0.45 (0.08) | -0.58 (0.20) | 3.19 (0.55) | 5.02 (0.84) |
| 27 | 0.79 (0.08) | -0.92 (0.13) | 1.59 (0.18) | 3.24 (0.34) |
| 29 | 0.63 (0.08) | -0.84 (0.16) | 2.11 (0.27) | 3.77 (0.47) |

Nota: los errores estándar aparecen entre paréntesis.

Antes de analizar la equivalencia de medida entre hombres y mujeres se igualan los parámetros de los ítems de la escala a una métrica común. Los parámetros del ítem del grupo focal fueron los que se igualaron a la métrica subyacente de los parámetros de los ítems del grupo de referencia. Se consideró como grupo de referencia al que contaba con un mayor número de sujetos, en este caso, el de mujeres. Los coeficientes de transformación métrica son, en este caso: A = 0.9151 y K = 0.1359.

Con estos coeficientes se transforman los parámetros de grupo de hombres en la métrica subyacente del grupo de mujeres. Los parámetros del ítem estimados para hombres y mujeres, una vez igualada la métrica pueden verse en la tabla 3.66, junto con los índices CDIF, NCDIF y el punto de corte asociado.

Tabla 3.66. *Parámetros estimados del ítem para hombres y mujeres, e índices de funcionamiento diferencial del ítem*

| Item | *a* | *b1* | *b2* | *b3* | *CDIF* | *NCDIF* | Pto. Corte* | Sig. |
|---|---|---|---|---|---|---|---|---|
| Item 1 | | | | | 0.00156 | 0.00163 | 0.00726 | ns |
| Hombres | 1.10 | -1.49 | 1.10 | 3.47 | | | | |
| Mujeres | 0.99 | -1.37 | 0.15 | 4.11 | | | | |
| Item 2 | | | | | -0.00049 | 0.00022 | 0.00281 | ns |
| Hombres | 1.49 | -0.74 | 1.46 | 2.98 | | | | |
| Mujeres | 1.49 | -0.79 | 1.43 | 3.04 | | | | |
| Item 3 | | | | | 0.00414 | 0.01289 | 0.00380 | .001 |
| Hombres | 1.14 | -0.91 | 1.35 | 2.89 | | | | |
| Mujeres | 1.19 | -0.67 | 1.76 | 3.23 | | | | |
| Item 4 | | | | | -0.00101 | 0.00059 | 0.00860 | ns |
| Hombres | 0.42 | -2.70 | 0.64 | 4.24 | | | | |
| Mujeres | 0.49 | -2.88 | 0.57 | 3.78 | | | | |
| Item 5 | | | | | 0.00038 | 0.00987 | 0.01012 | ns |
| Hombres | 0.57 | -2.72 | -0.32 | 2.50 | | | | |
| Mujeres | 0.38 | -4.31 | -0.57 | 4.34 | | | | |
| Item 6 | | | | | -0.00122 | 0.00110 | 0.00859 | ns |
| Hombres | 0.45 | 1.28 | 3.23 | 4.41 | | | | |
| Mujeres | 0.46 | 0.84 | 3.16 | 4.29 | | | | |
| Item 7 | | | | | -0.00278 | 0.00467 | 0.00800 | ns |
| Hombres | 1.13 | -1.38 | -0.11 | 2.06 | | | | |
| Mujeres | 1.29 | -1.40 | -0.28 | 1.69 | | | | |
| Item 8 | | | | | -0.00483 | 0.01495 | 0.01011 | .001 |
| Hombres | 0.55 | -1.64 | 0.61 | 2.76 | | | | |
| Mujeres | 0.66 | -1.74 | -0.04 | 2 | | | | |
| Item 9 | | | | | 0.00110 | 0.00071 | 0.00810 | ns |
| Hombres | 0.79 | -0.87 | 1.00 | 2.30 | | | | |
| Mujeres | 0.72 | -0.75 | 1.19 | 2.47 | | | | |
| Item 10 | | | | | 0.00137 | 0.00188 | 0.00634 | ns |
| Hombres | 0.87 | -1.26 | 0.57 | 3.21 | | | | |
| Mujeres | 0.89 | -1.13 | 0.63 | 3.3 | | | | |
| Item 11 | | | | | 0.00269 | 0.00475 | 0.00369 | .005 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Hombres | 1.05 | 0.12 | 1.45 | 3.25 | | | | |
| Mujeres | 1.04 | 0.35 | 1.69 | 3.51 | | | | |
| Item 12 | | | | | 0.00013 | 0.00047 | 0.00332 | ns |
| Hombres | 1.41 | -0.70 | 1.23 | 2.42 | | | | |
| Mujeres | 1.63 | -0.72 | 1.16 | 2.34 | | | | |
| Item 13 | | | | | -0.00561 | 0.02935 | 0.00913 | .001 |
| Hombres | 0.36 | -4.01 | -0.61 | 3.73 | | | | |
| Mujeres | 0.39 | -4.87 | 1.42 | 2.92 | | | | |
| Item 14 | | | | | 0.00162 | 0.00161 | 0.00318 | ns |
| Hombres | 1.23 | 0.21 | 1.36 | 2.42 | | | | |
| Mujeres | 1.12 | 0.22 | 1.80 | 2.89 | | | | |
| Item 15 | | | | | 0.00044 | 0.00237 | 0.00392 | ns |
| Hombres | 1.17 | -0.48 | 1.29 | 2.56 | | | | |
| Mujeres | 1.37 | -0.39 | 1.06 | 2.21 | | | | |
| Item 16 | | | | | -0.00215 | 0.00427 | 0.00516 | ns |
| Hombres | 1.09 | -0.6 | 1.08 | 2.01 | | | | |
| Mujeres | 1.16 | -0.92 | 1.03 | 2.09 | | | | |
| Item 17 | | | | | 0.00304 | 0.00771 | 0.01152 | ns |
| Hombres | 0.53 | -0.72 | 0.55 | 2.35 | | | | |
| Mujeres | 0.33 | -1.17 | 1.49 | 5.46 | | | | |
| Item 18 | | | | | -0.00035 | 0.00032 | 0.00280 | ns |
| Hombres | 1.76 | -0.34 | 1.27 | 2.35 | | | | |
| Mujeres | 1.78 | -0.41 | 1.40 | 2.34 | | | | |
| Item 19 | | | | | 0.00004 | 0.00061 | 0.00652 | ns |
| Hombres | 1.36 | -1.18 | 0.19 | 1.97 | | | | |
| Mujeres | 1.21 | -1.17 | 0.16 | 2.22 | | | | |
| Item 20 | | | | | 0.00043 | 0.00013 | 0.00284 | ns |
| Hombres | 1.26 | -0.20 | 1.70 | 2.82 | | | | |
| Mujeres | 1.29 | -0.20 | 1.77 | 2.83 | | | | |
| Item 21 | | | | | -0.00163 | 0.00162 | 0.00507 | ns |
| Hombres | 0.78 | -0.11 | 1.66 | 3.33 | | | | |
| Mujeres | 0.93 | -0.40 | 1.42 | 2.81 | | | | |
| Item 22 | | | | | -0.00082 | 0.00097 | 0.00424 | ns |
| Hombres | 1.06 | -0.12 | 1.30 | 3.23 | | | | |
| Mujeres | 1.21 | -0.13 | 0.97 | 2.93 | | | | |
| Item 24 | | | | | 0.00227 | 0.00302 | 0.00350 | ns |
| Hombres | 1.03 | -0.03 | 1.63 | 2.97 | | | | |
| Mujeres | 0.97 | -0.06 | 2.07 | 3.69 | | | | |
| Item 25 | | | | | 0.00587 | 0.02938 | 0.00199 | .001 |
| Hombres | 0.56 | 1.51 | 3.66 | 5.19 | | | | |
| Mujeres | 0.85 | 1.57 | 3.48 | 4.65 | | | | |
| Item 26 | | | | | -0.00103 | 0.00210 | 0.00761 | ns |
| Hombres | 0.29 | -0.71 | 5.16 | 8.95 | | | | |
| Mujeres | 0.45 | -0.59 | 3.19 | 5.02 | | | | |
| Item 27 | | | | | 0.00378 | 0.01461 | 0.00567 | .001 |
| Hombres | 0.66 | -1.12 | 1.34 | 3.09 | | | | |
| Mujeres | 0.79 | -0.92 | 1.59 | 3.24 | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Item 29 | | | | | -0.00421 | 0.01174 | 0.00621 | .001 |
| Hombres | 0.52 | -0.17 | 2.82 | 5.02 | | | | |
| Mujeres | 0.64 | -0.84 | 2.11 | 3.77 | | | | |

*Punto de corte establecido del índice NCDIF con el procedimiento IPR para α =.01

La suma de los valores del índice CDIF de todos los ítems del test proporciona el valor del índice DTF, que es 0.00197. Este valor de funcionamiento diferencial es muy inferior al punto de corte, establecido en 0.16345, por lo que no hay indicios de funcionamiento diferencial del test entre hombres y mujeres. La CCT para hombres y mujeres (ver Figura 3.34) refleja esta situación de equivalencia entre ambos sexos, al estar ambas líneas solapadas para todos los niveles de impulsividad.

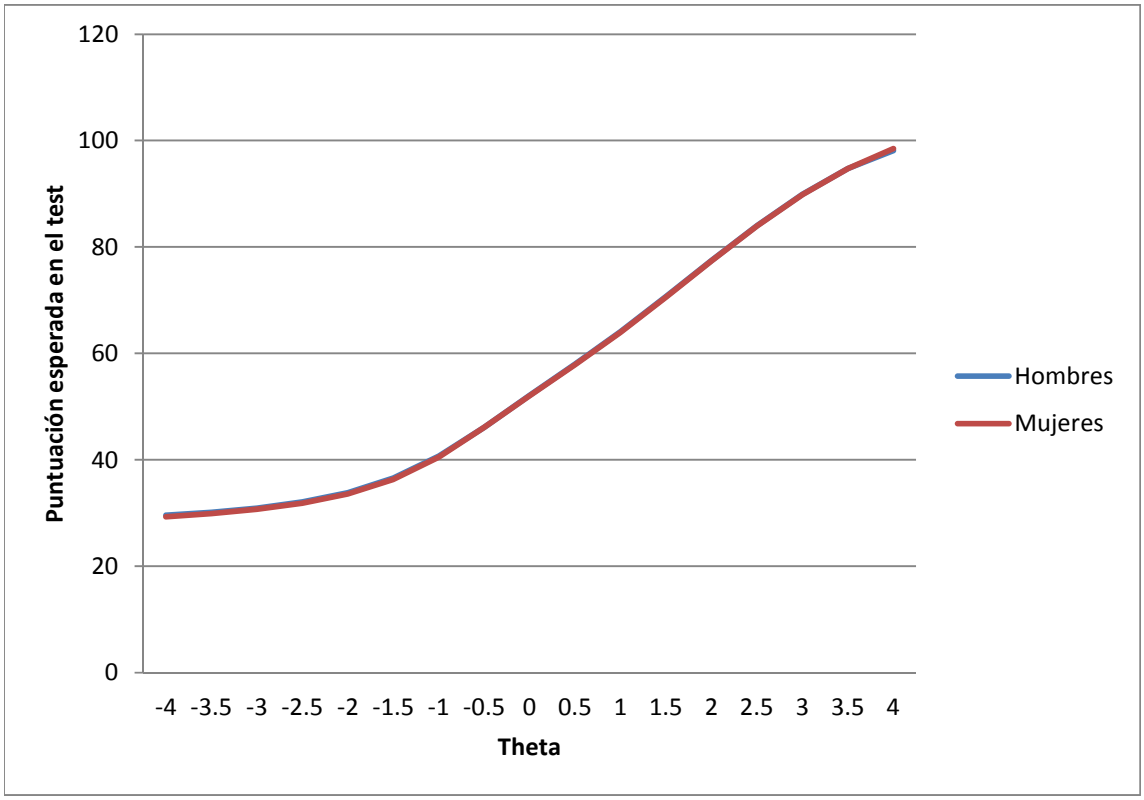



Figura 3.34 CCT en hombres y mujeres en el test completo BIS.

A nivel de ítem, encontramos que, según el índice no compensatorio NCDIF, 4 ítems presentan DIF en una dirección y 3 en la opuesta. En total son 7 los ítems con

funcionamiento diferencial utilizando el punto de corte propuesto para el índice no compensatorio NCDIF: los ítems 3, 8, 11, 13, 25, 27 y 29.

Estos resultados son algo diferentes de los encontrados analizando cada subescala por separado, habiendo un menor número de ítems con DIF en el caso de la escala completa:

(1) En la subescala Impulso Motor coinciden ambos resultados, siendo detectado únicamente el ítem 29.

(2) En la subescala Impulso No Planificado en ambos casos se detectan 4 ítems, pero hay uno que no es coincidente: el ítem 11 presenta DIF en el análisis de la escala completa pero no en el de las subescalas, y el ítem 17 es el caso contrario, ya que presenta DIF en el análisis de la subescala pero no en el de la escala completa.

(3) En la subescala Impulso Cognitivo-Atencional es donde se encuentran más diferencias, habiendo únicamente dos ítems con DIF en el análisis de la escala completa (13 y 27), y 5 en el análisis de la subescala (7, 10, 13, 24 y 27).

## 3.5.2. EQUIVALENCIA DE MEDIDA ENTRE PREADOLESCENTES Y ADOLESCENTES

### *3.5.2.1. Subescala Impulso Motor del BIS*

En las tablas 3.67 y 3.68 se muestra la estimación de los parámetros de los ítems de la subescala Impulso Motor del BIS-PA para preadolescentes y adolescentes.

Tabla 3.67. *Parámetros estimados (y errores estándar asociados) para los ítems de la subescala Impulso Motor del BIS en la muestra de preadolescentes*

| Item | *a* | *b1* | *b2* | *b3* |
|---|---|---|---|---|
| 2 | 1.32 (0.10) | -0.37 (0.07) | 1.98 (0.15) | 3.37 (0.28) |
| 6 | 0.42 (0.08) | 1.11 (0.29) | 3.39 (0.70) | 4.45 (0.92) |
| 9 | 0.67 (0.08) | -0.74 (0.15) | 1.55 (0.21) | 2.94 (0.36) |
| 12 | 1.88 (0.11) | -0.33 (0.06) | 1.33 (0.08) | 2.24 (0.13) |
| 15 | 1.31 (0.10) | -0.14 (0.07) | 1.49 (0.11) | 2.63 (0.20) |
| 18 | 1.81 (0.12) | -0.03 (0.06) | 1.56 (0.10) | 2.33 (0.15) |
| 26 | 0.42 (0.08) | -0.08 (0.20) | 4.32 (0.81) | 6.79 (1.28) |
| 29 | 0.69 (0.08) | -0.46 (0.13) | 2.09 (0.26) | 3.57 (0.43) |

Nota: los errores estándar aparecen entre paréntesis.

Tabla 3.68. *Parámetros estimados (y errores estándar asociados) para los ítems de la subescala Impulso Motor del BIS en la muestra de adolescentes*

| Item | *a* | *b1* | *b2* | *b3* |
|---|---|---|---|---|
| 2 | 2.38 (0.13) | -1.06 (0.06) | 0.93 (0.06) | 2.35 (0.13) |
| 6 | 0.35 (0.07) | 1.38 (0.40) | 4.24 (1.04) | 5.83 (1.43) |
| 9 | 0.74 (0.07) | -1.04 (0.16) | 0.80 (0.14) | 2.12 (0.25) |
| 12 | 2.25 (0.12) | -0.99 (0.06) | 0.79 (0.06) | 1.92 (0.10) |
| 15 | 1.17 (0.08) | -0.82 (0.10) | 0.93 (0.10) | 2.27 (0.18) |
| 18 | 2.33 (0.12) | -0.73 (0.06) | 1.02 (0.06) | 2.15 (0.12) |
| 26 | 0.35 (0.07) | -1.10 (0.34) | 3.49(0.79) | 6.05 (1.35) |
| 29 | 0.48 (0.07) | -0.69 (0.21) | 2.87 (0.48) | 5.25 (0.87) |

Nota: los errores estándar aparecen entre paréntesis.

Para analizar la equivalencia de medida entre ambos grupos es necesario igualar los parámetros de los ítems del grupo focal a la métrica subyacente de los parámetros de los ítems del grupo de referencia. Dado que el grupo de adolescentes es más numeroso se le considera el grupo de referencia, siendo los preadolescentes el grupo focal.

Los coeficientes de transformación métrica son: A = 1.0592 y K = -0.5454. Una vez que ambos grupos cuentan con una métrica común se ha comparado la equivalencia de medida entre ambos grupos.

Para valorar si existe funcionamiento diferencial de la subescala hay que comparar el valor del índice DTF con un punto de corte establecido según las características de los ítems de la subescala por el procedimiento IPR. En este caso este valor es 0.04794. El valor de DTF en la subescala Impulso Motor es 0.02563. Puesto que este valor es menor que 0.04794, se puede afirmar que hay equivalencia entre preadolescentes y adolescentes en esta subescala. En la Figura 3.35 se representa gráficamente la CCT para ambos grupos de edad, que es consistente con este resultado de equivalencia, ya que las líneas de ambos grupos de edad están solapadas en todo el continuo excepto en niveles muy altos de impulsividad.

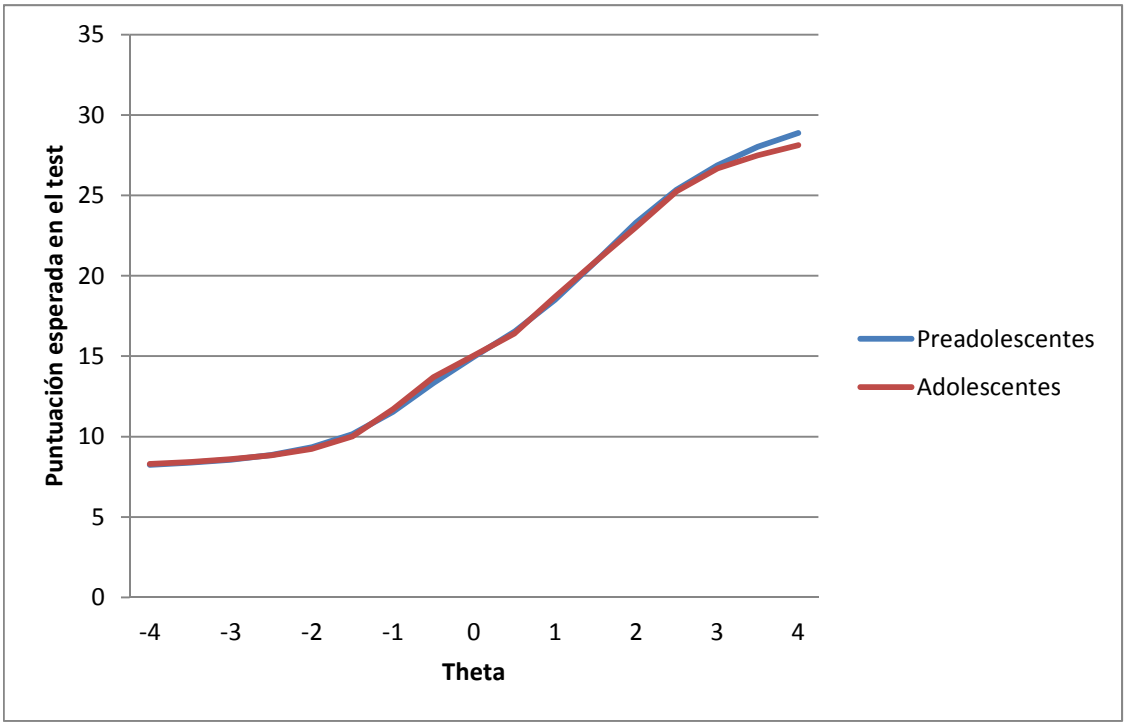



Figura 3.35. CCT para preadolescentes y adolescentes en la subescala Impulso Motor del BIS.

En cuanto al funcionamiento diferencial a nivel de ítem, el valor de NCDIF es significativo en tres de los ocho ítems de la escala: el ítem 2, el ítem 26 y el ítem 29 (ver

tabla 3.69) por lo que presentan DIF en relación con la edad. La dirección del DIF del ítem 29 es opuesta a las de los ítems 2 y 26 (ver columna CDIF de la tabla 3.69), por lo que ha habido una cierta compensación a nivel de subescala del funcionamiento diferencial.

Tabla 3.69. *Parámetros estimados del ítem para preadolescentes y adolescentes en la subescala Impulso Motor, e índices de funcionamiento diferencial del ítem*

| Item | *a* | *b1* | *b2* | *b3* | *CDIF* | *NCDIF* | Pto. Corte* | Sig. |
|---|---|---|---|---|---|---|---|---|
| Item 2 | | | | | 0.01682 | 0.01273 | 0.00244 | .001 |
| Preadolescentes | 1.25 | -0.94 | 1.55 | 3.02 | | | | |
| Adolescentes | 2.38 | -1.06 | 0.93 | 2.35 | | | | |
| Item 6 | | | | | -0.01418 | 0.01221 | 0.01306 | ns |
| Preadolescentes | 0.40 | 0.63 | 3.05 | 4.17 | | | | |
| Adolescentes | 0.35 | 1.39 | 4.24 | 5.83 | | | | |
| Item 9 | | | | | 0.00003 | 0.00213 | 0.00863 | ns |
| Preadolescentes | 0.63 | -1.33 | 1.09 | 2.57 | | | | |
| Adolescentes | 0.74 | -1.04 | 0.80 | 2.12 | | | | |
| Item 12 | | | | | 0.00514 | 0.00110 | 0.00271 | ns |
| Preadolescentes | 1.78 | -0.90 | 0.86 | 1.83 | | | | |
| Adolescentes | 2.26 | -0.99 | 0.80 | 1.92 | | | | |
| Item 15 | | | | | 0.00960 | 0.00603 | 0.00654 | ns |
| Preadolescentes | 1.23 | -0.69 | 1.03 | 2.24 | | | | |
| Adolescentes | 1.17 | -0.82 | 0.93 | 2.27 | | | | |
| Item 18 | | | | | 0.00675 | 0.00197 | 0.00243 | ns |
| Preadolescentes | 1.71 | -0.51 | 1.11 | 1.92 | | | | |
| Adolescentes | 2.33 | -0.73 | 1.02 | 2.15 | | | | |
| Item 26 | | | | | 0.01725 | 0.02058 | 0.00770 | .001 |
| Preadolescentes | 0.4 | -0.63 | 4.03 | 6.64 | | | | |
| Adolescentes | 0.35 | -1.10 | 3.49 | 6.05 | | | | |
| Item 29 | | | | | -0.01579 | 0.01404 | 0.00695 | .001 |
| Preadolescentes | 0.65 | -1.04 | 1.67 | 3.24 | | | | |
| Adolescentes | 0.48 | -0.69 | 2.87 | 5.25 | | | | |

*Punto de corte establecido del índice NCDIF con el procedimiento IPR para α =.01

### *3.5.2.2. Subescala Impulso no Planificado del BIS*

A continuación se muestran los parámetros de cada ítem de la subescala BIS-PA para preadolescentes y adolescentes por separado (ver Tablas 3.70 y 3.71).

Tabla 3.70. *Parámetros estimados (y errores estándar asociados) para los ítems de la subescala Impulso No Planificado del BIS en la muestra de preadolescentes*

| Item | *a* | *b1* | *b2* | *b3* |
|---|---|---|---|---|
| 1 | 0.70 (0.08) | -1.20 (0.17) | 0.73 (0.14) | 5.52 (0.69) |
| 3 | 0.83 (0.09) | -0.80 (0.12) | 2.32 (0.23) | 3.88 (0.40) |
| 5 | 0.31 (0.07) | -4.02 (0.96) | -0.32 (0.27) | 4.64 (1.11) |
| 8 | 0.40 (0.07) | -1.53 (0.33) | 1.19 (0.29) | 4.24 (0.80) |
| 11 | 1.12 (0.11) | 0.90 (0.10) | 2.07 (0.18) | 3.27 (0.31) |
| 14 | 1.37 (0.11) | 0.68 (0.07) | 1.87 (0.14) | 2.47 (0.19) |
| 17 | 0.52 (0.08) | 0.05 (0.15) | 1.66 (0.28) | 3.76 (0.56) |
| 22 | 1.63 (0.12) | 0.35 (0.06) | 1.11 (0.08) | 2.32 (0.15) |
| 25 | 0.62 (0.10) | 2.20 (0.35) | 4.05 (0.65) | 5.07 (0.82) |

Nota: los errores estándar aparecen entre paréntesis.

Tabla 3.71. *Parámetros estimados (y errores estándar asociados) para los ítems de la subescala Impulso No Planificado del BIS en la muestra de adolescentes*

| Item | *a* | *b1* | *b2* | *b3* |
|---|---|---|---|---|
| 1 | 1.08 (0.09) | -2.10 (0.18) | -0.30 (0.08) | 3.48 (0.31) |
| 3 | 1.23 (0.09) | -0.93 (0.09) | 1.28 (0.11) | 2.98 (0.23) |
| 5 | 0.54 (0.07) | -3.66 (0.49) | -0.64 (0.18) | 2.96 (0.43) |
| 8 | 0.62 (0.07) | -2.38 (0.32) | -0.29 (0.14) | 1.88 (0.25) |
| 11 | 1.28 (0.10) | -0.34 (0.08) | 1.01 (0.09) | 2.98 (0.23) |
| 14 | 1.23 (0.10) | -0.29 (0.08) | 1.20 (0.10) | 2.57 (0.19) |
| 17 | 0.33 (0.07) | -2.39 (0.57) | -0.13 ()0.24 | 3.43 (0.76) |
| 22 | 1.50 (0.10) | -0.55 (0.07) | 0.78 (0.07) | 2.81 (0.20) |
| 25 | 0.49 (0.08) | 1.67 (0.32) | 4.63 (0.82) | 6.90 (1.24) |

Nota: los errores estándar aparecen entre paréntesis.

Es necesario que los datos de ambos grupos posean una métrica común. Los parámetros de los ítems del grupo de preadolescentes (grupo focal) se igualan a la métrica del los del grupo de adolescentes (grupo de referencia), basándonos en los siguientes coeficientes de transformación: A = 1.0145 y K = -0.8467.

En la siguiente tabla (ver Tabla 3.72) aparecen los parámetros de los ítems de la subescala, ya en una métrica común, para ambos grupos, así como los índices de funcionamiento diferencial, basados en el procedimiento DFIT, para cada ítem de la subescala Impulso No Planificado.

Tabla 3.72. *Parámetros estimados del ítem para preadolescentes y adolescentes en la subescala Impulso No Planificado, e índices de funcionamiento diferencial del ítem*

| Item | *a* | *b1* | *b2* | *b3* | *CDIF* | *NCDIF* | Pto. Corte* | Sig. |
|---|---|---|---|---|---|---|---|---|
| Item 1 | | | | | 0.02694 | 0.01348 | 0.00458 | .001 |
| Preadolescentes | 0.69 | -2.06 | -0.11 | 4.76 | | | | |
| Adolescentes | 1.08 | -2.10 | -0.30 | 3.48 | | | | |
| Item 3 | | | | | -0.04238 | 0.03359 | 0.00378 | .001 |
| Preadolescentes | 0.82 | -1.66 | 1.50 | 3.09 | | | | |
| Adolescentes | 1.23 | -0.93 | 1.28 | 2.98 | | | | |
| Item 5 | | | | | -0.01040 | 0.00434 | 0.00772 | ns |
| Preadolescentes | 0.31 | -4.93 | -1.17 | 3.86 | | | | |
| Adolescentes | 0.54 | -3.66 | -0.64 | 2.96 | | | | |
| Item 8 | | | | | -0.04127 | 0.03135 | 0.00949 | .001 |
| Preadolescentes | 0.40 | -2040 | 0.36 | 3.46 | | | | |
| Adolescentes | 0.62 | -2.38 | -0.29 | 1.88 | | | | |
| Item 11 | | | | | 0.01950 | 0.00636 | 0.00514 | .005 |
| Preadolescentes | 1.10 | 0.07 | 1.25 | 2.47 | | | | |
| Adolescentes | 1.28 | -0.34 | 1.01 | 2.98 | | | | |
| Item 14 | | | | | -0.00049 | 0.00245 | 0.00471 | ns |
| Preadolescentes | 1.35 | -0.16 | 1.06 | 1.66 | | | | |
| Adolescentes | 1.23 | -0.29 | 1.20 | 2.57 | | | | |
| Item 17 | | | | | 0.07303 | 0.09355 | 0.01215 | .001 |
| Preadolescentes | 0.52 | -0.80 | 0.84 | 2.97 | | | | |
| Adolescentes | 0.34 | -2.39 | -0.13 | 3.43 | | | | |
| Item 22 | | | | | -0.03489 | 0.02891 | 0.000446 | .001 |
| Preadolescentes | 1.61 | -0.49 | 0.28 | 1.51 | | | | |
| Adolescentes | 1.50 | -0.55 | 0.78 | 2.81 | | | | |
| Item 25 | | | | | -0.00775 | 0.00213 | 0.00541 | ns |
| Preadolescentes | 0.61 | 1.39 | 3.27 | 4.29 | | | | |
| Adolescentes | 0.49 | 1.67 | 4.63 | 6.90 | | | | |

*Punto de corte establecido del índice NCDIF con el procedimiento IPR para α =.01

De los 9 ítems de la subescala, 6 presentan funcionamiento diferencial: los ítems 1, 3, 8, 11, 17 y 22, al ser su valor de NCDIF mayor que el punto de corte establecido para ese ítem (ver tabla 3.72).

La CCT para ambos grupos de edad (Figura 3.36) muestra que en niveles intermedios de theta la puntuación esperada en el test es mayor para los adolescentes, invirtiéndose en los niveles altos de impulsividad no planificada, en los que se aprecia una puntuación esperada en el test mayor para los preadolescentes.

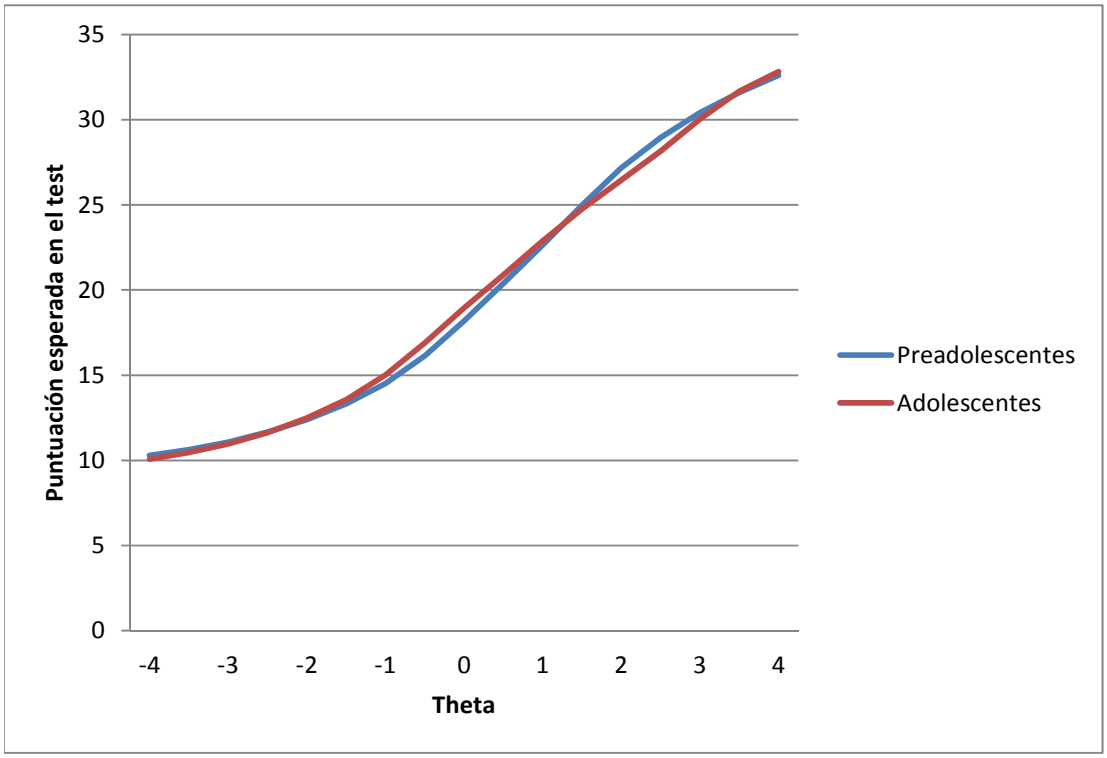



Figura 3.36. CCT para preadolescentes y adolescentes en la subescala Impulso No Planificado del BIS.

A nivel de subescala se considera que existe funcionamiento diferencial entre ambos grupos de edad si el valor del índice DTF es mayor que 0.05794. Puesto que DTF =

0.06483, hay funcionamiento diferencial a nivel de subescala, siendo necesario eliminar el ítem 17 de la subescala para obtener un valor de DTF no significativo (0.01233). En la tabla 3.73 aparecen los datos de este proceso iterativo.

Tabla 3.73. *Procedimiento iterativo de eliminación de ítems para establecer la equivalencia de medida en la subescala Impulso No Planificado*

| **Nº Ejecución** | **Ítem eliminado** | **DTF** | **Pto corte DTF** | **Sig** |
|---|---|---|---|---|
| **1** | Ninguno | 0.06484 | 0.05794 | .01 |
| **2** | Ítem 17 | 0.01233 | 0.05794 | ns |

Dado que es necesario eliminar uno de los ítems con DIF para establecer la equivalencia de medida entre ambos grupos de edad es aconsejable eliminarlo y volver a estimar las constantes de igualación en una métrica común, por la posible influencia que haya podido tener el sesgo de este ítem en el procedimiento de igualación. Las nuevas constantes de igualación son: A = 0.958 y K = -0.689. Con ellas, se re-estiman los parámetros de cada ítem para cada grupo por separado y se calculan los índices de funcionamiento diferencial (ver tabla 3.74).

Tabla 3.74. *Parámetros estimados del ítem para preadolescentes (2ª equiparación) y adolescentes en la subescala Impulso No Planificado, e índices de funcionamiento diferencial del ítem*

| Item | *a* | *b1* | *b2* | *b3* | *CDIF* | *NCDIF* | Pto. Corte* | Sig. |
|---|---|---|---|---|---|---|---|---|
| Item 1 | | | | | 0.09792 | 0.02325 | 0.00450 | .001 |
| Preadolescentes | 0.73 | -1.85 | -0.00 | 4.61 | | | | |
| Adolescentes | 1.09 | -2.10 | -0.30 | 3.49 | | | | |
| Item 3 | | | | | 0.01963 | 0.01963 | 0.00418 | .001 |
| Preadolescentes | 0.87 | -1.47 | 1.53 | 3.03 | | | | |
| Adolescentes | 1.21 | -0.94 | 1.29 | 3.02 | | | | |
| Item 5 | | | | | 0.00217 | 0.00217 | 0.00725 | ns |
| Preadolescentes | 0.32 | -4.57 | -1.00 | 3.76 | | | | |
| Adolescentes | 0.53 | -3.78 | -0.64 | 3.03 | | | | |
| Item 8 | | | | | 0.04224 | 0.04224 | 0.00995 | .001 |
| Preadolescentes | 0.42 | -2.17 | 0.44 | 3.38 | | | | |
| Adolescentes | 0.63 | -2.39 | -0.29 | 1.82 | | | | |
| Item 11 | | | | | 0.01749 | 0.01749 | 0.00493 | .001 |
| Preadolescentes | 1.16 | 0.17 | 1.29 | 2.44 | | | | |
| Adolescentes | 1.27 | -0.34 | 1.01 | 2.98 | | | | |
| Item 14 | | | | | 0.00612 | 0.00612 | 0.00486 | .005 |
| Preadolescentes | 1.43 | -0.04 | 1.10 | 1.68 | | | | |
| Adolescentes | 1.23 | -0.29 | 1.20 | 2.57 | | | | |
| Item 17 | | | | | 0.12113 | 0.12113 | 0.01330 | .001 |
| Preadolescentes | 0.54 | 0.66 | 0.90 | 2.91 | | | | |
| Adolescentes | 0.34 | -2.35 | -0.12 | 3.35 | | | | |
| Item 22 | | | | | 0.01203 | 0.01203 | 0.00449 | .001 |
| Preadolescentes | 1.70 | -0.36 | 0.36 | 1.53 | | | | |
| Adolescentes | 1.54 | -0.55 | 0.76 | 2.77 | | | | |
| Item 25 | | | | | 0.00122 | 0.00122 | 0.00572 | ns |
| Preadolescentes | 0.64 | 1.42 | 3.20 | 4.17 | | | | |
| Adolescentes | 0.47 | 1.69 | 4.72 | 7.16 | | | | |

*Punto de corte establecido del índice NCDIF con el procedimiento IPR para α =.01

Ahora la CCT para ambos grupos de edad (Figura 3.37) muestra un mayor desajuste en los niveles intermedios de impulso no planificado, en los que la puntuación esperada en el test es mayor para adolescentes, invirtiéndose ligeramente en los niveles altos del rasgo, en los que se aprecia una puntuación esperada en el test mayor para los preadolescentes.

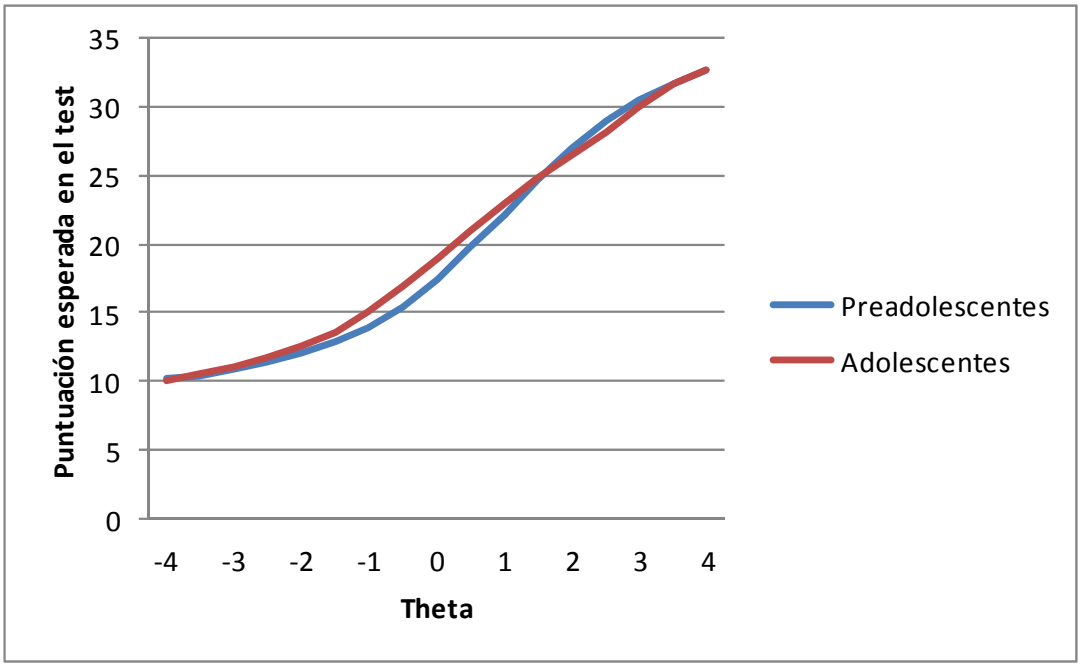



Figura 3.37. CCT para preadolescentes y adolescentes en la subescala Impulso No Planificado del BIS tras la segunda equiparación.

En la subescala completa Impulso No Planificado, el valor de DTF arroja un resultado de 0.45042, muy superior al punto de corte establecido (<0.06067), por lo que la subescala presenta funcionamiento diferencial. Eliminando el ítem 17 el procedimiento iterativo arroja un valor de DTF igual a 0.11367, que sigue siendo superior al punto de corte, por tanto indicativo de funcionamiento diferencial a nivel de la escala. Es necesario eliminar otro ítem más para obtener la equivalencia en la subescala, el ítem 8, obteniendo así un valor de 0.02439 para el DTF, que no es significativo (ver tabla 3.75).

Tabla 3.75. *Segundo procedimiento iterativo de eliminación de ítems para establecer la equivalencia de medida en la subescala Impulso No Planificado*

| Nº Ejecución | Ítem eliminado | DTF | Pto corte DTF | Sig |
|---|---|---|---|---|
| **1** | Ninguno | 0.45042 | 0.06067 | .001 |
| **2** | Ítem 17 | 0.11367 | 0.06067 | .001 |
| **3** | Ítem 8 | 0.02439 | 0.06067 | ns |

### *3.5.2.3. Subescala Impulso Cognitivo-Atencional del BIS*

En las tablas 3.76 y 3.77 se muestran los parámetros de los ítems de la subescala ICA para preadolescentes y adolescentes.

Tabla 3.76. *Parámetros estimados (y errores estándar asociados) para los ítems de la subescala Impulso Cognitivo-Atencional del BIS en la muestra de preadolescentes*

| Item | *a* | *b1* | *b2* | *b3* |
|---|---|---|---|---|
| 4 | 0.22 (0.08) | -3.43 (1.25) | 3.08 (1.09) | 8.54 (2.94) |
| 7 | 1.32 (0.09) | -0.98 (0.09) | 0.13 (0.07) | 2.20 (0.16) |
| 10 | 1.13 (0.09) | -0.98 (0.11) | 0.38 (0.08) | 2.58 (0.20) |
| 13 | 0.62 (0.07) | -2.17 (0.30) | -0.60 (0.15) | 1.76 (0.24) |
| 16 | 1.27 (0.10) | -0.35 (0.08) | 1.20 (0.11) | 2.00 (0.16) |
| 19 | 1.20 (0.10) | -0.95 (0.10) | 0.35 (0.08) | 2.27 (0.18) |
| 20 | 1.11 (0.10) | 0.26 (0.08) | 2.18 (0.18) | 3.18 (0.28) |
| 21 | 0.58 (0.08) | 0.46 (0.15) | 2.94 (0.42) | 4.69 (0.68) |
| 24 | 0.92 (0.09) | 0.51 (0.10) | 2.16 (0.22) | 3.26 (0.34) |
| 27 | 0.41 (0.07) | -1.22 (0.29) | 2.91 (0.57) | 5.43 (1.04) |

Nota: los errores estándar aparecen entre paréntesis.

Tabla 3.77. *Parámetros estimados (y errores estándar asociados) para los ítems de la subescala Impulso Cognitivo-Atencional del BIS en la muestra de adolescentes*

| Item | *a* | *b1* | *b2* | *b3* |
|---|---|---|---|---|
| 4 | 0.30 (0.07) | -6.43 (1.55) | -0.32 (0.59) | 5.63 (1.35) |
| 7 | 1.24 (0.09) | -1.85 (0.14) | -0.56 (0.08) | 1.50 (0.12) |
| 10 | 0.99 (0.09) | -1.22 (0.14) | 0.64 (0.10) | 3.17 (0.29) |
| 13 | 0.53 (0.08) | -4.22 (0.64) | -0.96 (0.21) | 2.68 (0.40) |
| 16 | 1.30 (0.10) | -1.18 (0.10) | 0.72 (0.08) | 1.74 (0.14) |
| 19 | 1.12 (0.09) | -1.65 (0.15) | 0.00 (0.08) | 2.22 (0.19) |
| 20 | 1.12 (0.09) | -0.76 (0.09) | 1.73 (0.15) | 3.11 (0.27) |
| 21 | 0.54 (0.08) | -1.28 (0.24) | 1.66 (0.27) | 4.31 (0.63) |
| 24 | 0.47 (0.09) | -1.95 (0.11) | 1.89 (0.21) | 4.60 (0.46) |
| 27 | 0.83 (0.08) | -0.51 (0.36) | 1.90 (0.34) | 4.26 (0.74) |

Nota: los errores estándar aparecen entre paréntesis.

Para analizar el funcionamiento diferencial del ítem con el procedimiento DFIT, previamente se igualan los parámetros del grupo focal (preadolescentes) al grupo de referencia (adolescentes), calculando los coeficientes de transformación métrica: A = 1.1313 y K =-0.7233. En la siguiente tabla están los valores de los parámetros de los ítems de la subescala ICA una vez equiparada la métrica, así como diversos ítems del funcionamiento diferencial mediante el procedimiento DFIT.

Tabla 3.78. *Parámetros estimados del ítem para preadolescentes y adolescentes en la subescala Impulso Cognitivo-Atencional, e índices de funcionamiento diferencial del ítem*

| Item | *a* | *b1* | *b2* | *b3* | *CDIF* | *NCDIF* | Pto. corte | Sig. |
|---|---|---|---|---|---|---|---|---|
| Item 4 | | | | | -0.02582 | 0.10281 | 0.00740 | .001 |
| Preadolescentes | 0.20 | -4.60 | 2.72 | 8.84 | | | | |
| Adolescentes | 0.30 | -6.43 | -0.32 | 5.63 | | | | |
| Item 7 | | | | | -0.00206 | 0.00070 | 0.00552 | ns |
| Preadolescentes | 1.18 | -1.83 | -0.58 | 1.73 | | | | |
| Adolescentes | 1.24 | -1.85 | -0.56 | 1.50 | | | | |
| Item 10 | | | | | 0.02938 | 0.13428 | 0.00560 | .001 |
| Preadolescentes | 1.00 | -1.83 | -0.31 | 2.16 | | | | |
| Adolescentes | 0.99 | -1.22 | 0.64 | 3.17 | | | | |
| Item 13 | | | | | 0.01135 | 0.01946 | 0.00628 | .001 |
| Preadolescentes | 0.56 | -3.16 | -1.41 | 1.24 | | | | |
| Adolescentes | 0.53 | -4.22 | -0.96 | 2.68 | | | | |
| Item 16 | | | | | 0.00569 | 0.00488 | 0.00544 | ns |
| Preadolescentes | 1.13 | -1.13 | 0.61 | 1.51 | | | | |
| Adolescentes | 1.30 | -1.18 | 0.72 | 1.74 | | | | |
| Item 19 | | | | | 0.01060 | 0.01756 | 0.00458 | .001 |
| Preadolescentes | 1.07 | -1.80 | -0.34 | 1.81 | | | | |
| Adolescentes | 1.12 | -1.65 | 0.00 | 2.22 | | | | |
| Item 20 | | | | | -0.0329 | 0.00176 | 0.00365 | ns |
| Preadolescentes | 0.99 | -0.44 | 1.71 | 2.83 | | | | |
| Adolescentes | 1.12 | -0.76 | 1.73 | 3.11 | | | | |
| Item 21 | | | | | -0.01833 | 0.05215 | 0.00620 | .001 |
| Preadolescentes | 0.52 | -0.21 | 2.56 | 4.52 | | | | |
| Adolescentes | 0.54 | -1.28 | 1.66 | 4.31 | | | | |
| Item 24 | | | | | 0.00075 | 0.00068 | 0.00428 | ns |
| Preadolescentes | 0.81 | -0.16 | 1.68 | 2.91 | | | | |
| Adolescentes | 0.83 | -0.51 | 1.90 | 4.26 | | | | |
| Item 27 | | | | | -0.00163 | 0.00086 | 0.00766 | ns |
| Preadolescentes | 0.36 | -2.10 | 2.52 | 5.35 | | | | |
| Adolescentes | 0.47 | -1.95 | 1.89 | 4.60 | | | | |

*Punto de corte establecido del índice NCDIF con el procedimiento IPR para α =.01

El índice DTF, encargado de valorar el funcionamiento diferencial de la subescala arroja un resultado de 0.00664. Este valor se compara con 0.05017, considerado punto de corte para el global de la subescala. Por tanto, la subescala Impulso Cognitivo-Atencional no presenta funcionamiento diferencial. En la Figura 3.38 se representa gráficamente las

CCT para ambos grupos de edad, pudiéndose apreciar el solapamiento entre ambas líneas, lo que corrobora el resultado de equivalencia encontrado.

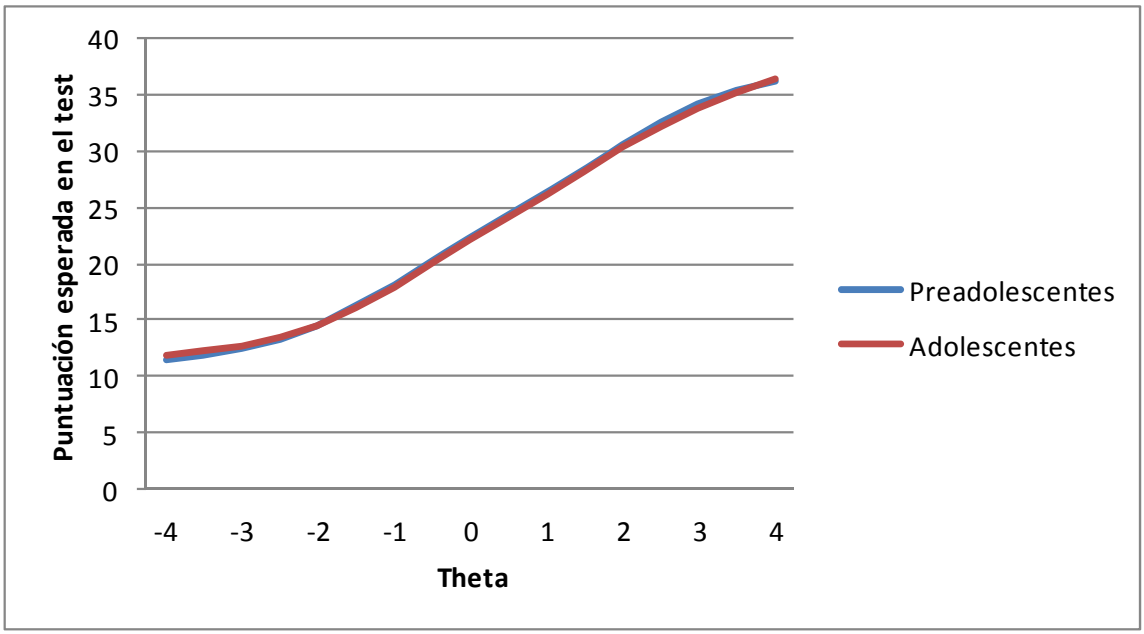



Figura 3.38. CCT para preadolescentes y adolescentes en la subescala Impulso Cognitivo-Atencional.

Sin embargo, a nivel de ítem encontramos cinco ítems con problemas de DIF (ver tabla 3.78): los ítems 4, 10, 13, 19 y 21.

### *3.5.2.4. Escala Total BIS*

Se estimaron los parámetros de los ítems de la escala para preadolescentes y adolescentes por separado (ver tablas 3.79 y 3.80).

Tabla 3.79. *Parámetros estimados (y errores estándar asociados) para los ítems del BIS en la muestra de preadolescentes*

| Item | *a* | *b1* | *b2* | *b3* |
|---|---|---|---|---|
| 1 | 0.73 (0.08) | -1.19 (0.18) | 0.66 (0.14) | 5.29 (0.65) |
| 2 | 1.04 (0.11) | -0.47 (0.09) | 2.29 (0.22) | 3.97 (0.41) |
| 3 | 1.03 (0.10) | -0.72 (0.10) | 1.93 (0.18) | 3.25 (0.30) |
| 4 | 0.29 (0.08) | -2.67 (0.79) | 2.35 (0.74) | 6.56 (2.07) |
| 5 | 0.36 (0.08) | -3.53 (0.85) | -0.31 (0.25) | 4.01 (0.85) |
| 6 | 0.48 (0.09) | 0.95 (0.24) | 2.96 (0.57) | 3.89 (0.76) |
| 7 | 1.08 (0.09) | -1.16 (0.12) | -0.11 (0.09) | 2.49 (0.21) |
| 8 | 0.31 (0.07) | -1.97 (0.55) | 1.48 (0.45) | 5.37 (1.29) |
| 9 | 0.68 (0.08) | -0.76 (0.15) | 1.48 (0.21) | 2.85 (0.37) |
| 10 | 0.98 (0.09) | -1.11 (0.13) | 0.39 (0.10) | 2.85 (0.26) |
| 11 | 0.85 (0.10) | 1.07 (0.15) | 2.51 (0.29) | 4.01 (0.50) |
| 12 | 1.27 (0.11) | -0.45 (0.08) | 1.60 (0.13) | 2.77 (0.23) |
| 13 | 0.42 (0.07) | -3.10 (0.62) | -0.87 (0.26) | 2.48 (0.47) |
| 14 | 1.17 (0.11) | 0.71 (0.09) | 2.05 (0.18) | 2.73 (0.25) |
| 15 | 1.10 (0.11) | -0.19 (0.08) | 1.62 (0.15) | 2.92 (0.26) |
| 16 | 1.21 (0.10) | -0.40 (0.08) | 1.18 (0.11) | 2.02 (0.17) |
| 17 | 0.41 (0.08) | 0.01 (0.21) | 2.04 (0.44) | 4.71 (0.91) |
| 18 | 1.52 (0.12) | -0.06 (0.07) | 1.67 (0.12) | 2.52 (0.19) |
| 19 | 1.25 (0.09) | -0.95 (0.10) | 0.32 (0.08) | 2.19 (0.18) |
| 20 | 1.15 (0.11) | 0.23 (0.08) | 2.10 (0.18) | 3.07 (0.28) |
| 21 | 0.77 (0.10) | 0.32 (0.12) | 2.27 (0.27) | 3.65 (0.45) |
| 22 | 1.06 (0.10) | 0.40 (0.09) | 1.38 (0.14) | 3.04 (0.29) |
| 24 | 0.88 (0.10) | 0.48 (0.11) | 2.19 (0.24) | 3.33 (0.37) |
| 25 | 0.72 (0.11) | 1.90 (0.29) | 3.52 (0.52) | 4.40 (0.68) |
| 26 | 0.27 (0.06) | -0.15 (0.26) | 6.50 (1.54) | 10.26 (2.43) |
| 27 | 0.67 (0.11) | -0.82 (0.15) | 1.82 (0.27) | 3.41 (0.49) |
| 29 | 0.70 (0.10) | -0.49 (0.14) | 2.03 (0.27) | 3.49 (0.46) |

Nota: los errores estándar aparecen entre paréntesis.

Tabla 3.80. *Parámetros estimados (y errores estándar asociados) para los ítems del BIS en la muestra de adolescentes*

| Item | *a* | *b1* | *b2* | *b3* |
|---|---|---|---|---|
| 1 | 1.09 (0.09) | -2.11 (0.18) | 0.32 (0.08) | 3.48 (0.32) |
| 2 | 1.76 (0.12) | -1.25 (0.08) | 0.98 (0.07) | 2.65 (0.17) |
| 3 | 1.28 (0.10) | -0.93 (0.09) | 1.21 (0.10) | 2.87 (0.23) |
| 4 | 0.34 (0.08) | -5.66 (1.29) | -0.35 (0.26) | 4.83 (1.11) |
| 5 | 0.51 (0.08) | -3.93 (0.63) | -0.70 (0.19) | 3.09 (0.48) |
| 6 | 0.41 (0.08) | 1.06 (0.29) | 3.51 (0.73) | 4.95 (1.02) |
| 7 | 0.98 (0.08) | -2.19 (0.21) | -0.67 (0.10) | 1.74 (0.16) |
| 8 | 0.71 (0.08) | -2.19 (0.27) | -0.30 (0.12) | 1.60 (0.21) |
| 9 | 0.80 (0.08) | -1.02 (0.15) | 0.71 (0.13) | 1.96 (0.22) |
| 10 | 1.02 (0.08) | -1.22 (0.13) | 0.57 (0.10) | 3.06 (0.26) |
| 11 | 0.88 (0.09) | -0.47 (0.11) | 1.24 (0.15) | 3.86 (0.40) |
| 12 | 1.59 (0.10) | -1.19 (0.09) | 0.86 (0.08) | 2.20 (0.14) |
| 13 | 0.35 (0.15) | -6.17 (2.66) | -1.41 (0.43) | 3.88 (1.63) |
| 14 | 0.91 (0.09) | -0.38 (0.10) | 1.44 (0.15) | 3.16 (0.30) |
| 15 | 1.22 (0.09) | -0.85 (0.10) | 0.84 (0.09) | 2.18 (0.18) |
| 16 | 0.86 (0.08) | -1.60 (0.18) | 0.95 (0.13) | 2.32 (0.23) |
| 17 | 0.22 (0.07) | -3.68 (1.34) | -0.21 (0.39) | 5.21 (1.78) |
| 18 | 1.78 (0.11) | -0.85 (0.07) | 1.09 (0.08) | 2.37 (0.16) |
| 19 | 1.33 (0.09) | -1.52 (0.12) | -0.05 (0.07) | 1.94 (0.14) |
| 20 | 1.10 (0.09) | -0.82 (0.10) | 1.68 (0.15) | 3.08 (0.26) |
| 21 | 0.66 (0.08) | -1.13 (0.18) | 1.33 (0.20) | 3.50 (0.44) |
| 22 | 0.96 (0.09) | -0.77 (0.11) | 0.99 (0.12) | 3.82 (0.37) |
| 24 | 0.88 (0.09) | -0.52 (0.11) | 1.78 (0.19) | 3.98 (0.41) |
| 25 | 0.57 (0.09) | 1.40 (0.25) | 3.97 (0.63) | 6.04 (0.99) |
| 26 | 0.29 (0.08) | -1.45 (0.48) | 4.11 (1.11) | 7.26 (1.92 |
| 27 | 0.66 (0.08) | -1.51 (0.21) | 1.31 (0.19) | 3.38 (0.41) |
| 29 | 0.45 (0.08) | -0.83 (0.24) | 2.92 (0.51) | 5.63 (0.97) |

Nota: los errores estándar aparecen entre paréntesis.

Antes de analizar la equivalencia de medida entre ambos grupos de edad se igualaron los parámetros de los ítems de la escala a una métrica común. Los coeficientes de transformación métrica son: A = 1.0788 y K = -0.7583. Con estos coeficientes se transforman los parámetros de los ítems del grupo de preadolescentes a la métrica del grupo de adolescentes (los valores obtenidos pueden verse en la tabla 3.81) y se calcula, con el procedimiento DFIT, un índice para el funcionamiento diferencial del test (DTF) y dos índices de DIF: NCDIF y CDIF.

Tabla 3.81. *Parámetros estimados del ítem para preadolescentes y adolescentes en el test BIS, e índices de funcionamiento diferencial del ítem*

| Item | *a* | *b1* | *b2* | *b3* | *CDIF* | *NCDIF* | Pto. corte | Sig. |
|---|---|---|---|---|---|---|---|---|
| Item 1 | | | | | 0.03700 | 0.01809 | 0.00456 | .001 |
| Preadolescentes | 0.68 | -2.05 | -0.50 | 4.94 | | | | |
| Adolescentes | 1.09 | -2.11 | -0.31 | 3.48 | | | | |
| Item 2 | | | | | 0.01858 | 0.00638 | 0.00262 | .001 |
| Preadolescentes | 0.96 | -1.27 | 1.71 | 3.53 | | | | |
| Adolescentes | 1.76 | -1.25 | 0.98 | 2.65 | | | | |
| Item 3 | | | | | -0.03954 | 0.02950 | 0.00406 | .001 |
| Preadolescentes | 0.96 | -1.54 | 1.33 | 2.75 | | | | |
| Adolescentes | 1.28 | -0.93 | 1.21 | 2.87 | | | | |
| Item 4 | | | | | 0.07503 | 0.08548 | 0.00789 | .001 |
| Preadolescentes | 0.27 | -3.64 | 1.78 | 6.32 | | | | |
| Adolescentes | 0.34 | -5.66 | -0.35 | 4.83 | | | | |
| Item 5 | | | | | -0.00531 | 0.00223 | 0.00645 | ns |
| Preadolescentes | 0.34 | -4.58 | -1.09 | 3.57 | | | | |
| Adolescentes | 0.51 | -3.93 | 0.70 | 3.09 | | | | |
| Item 6 | | | | | -0.04570 | 0.02994 | 0.01132 | .001 |
| Preadolescentes | 0.45 | 0.27 | 2.44 | 3.44 | | | | |
| Adolescentes | 0.41 | 1.06 | 3.51 | 4.95 | | | | |
| Item 7 | | | | | 0.01135 | 0.00218 | 0.00761 | ns |
| Preadolescentes | 1.00 | -2.00 | -0.65 | 1.93 | | | | |
| Adolescentes | 0.98 | -2.19 | -0.67 | 1.74 | | | | |
| Item 8 | | | | | 0.05792 | 0.04968 | 0.00865 | .001 |
| Preadolescentes | 0.29 | -2.88 | 0.84 | 5.03 | | | | |
| Adolescentes | 0.71 | -2.19 | -0.30 | 1.60 | | | | |
| Item 9 | | | | | -0.02238 | 0.01225 | 0.00828 | .005 |
| Preadolescentes | 0.63 | -1.58 | 0.84 | 2.32 | | | | |
| Adolescentes | 0.80 | -1.02 | 0.71 | 1.96 | | | | |
| Item 10 | | | | | -0.09310 | 0.13202 | 0.00545 | .001 |
| Preadolescentes | 0.90 | -1.95 | 0.34 | 2.31 | | | | |
| Adolescentes | 1.02 | -1.21 | -0.57 | 3.06 | | | | |
| Item 11 | | | | | 0.05254 | 0.03808 | 0.00613 | .001 |
| Preadolescentes | 0.79 | 0.39 | 1.95 | 3.57 | | | | |
| Adolescentes | 0.88 | -0.46 | 1.24 | 3.86 | | | | |
| Item 12 | | | | | -0.00667 | 0.00165 | 0.00354 | ns |
| Preadolescentes | 1.17 | -1.25 | 0.97 | 2.23 | | | | |
| Adolescentes | 1.59 | -1.19 | 0.86 | 2.19 | | | | |
| Item 13 | | | | | -0.01875 | 0.00488 | 0.00954 | ns |
| Preadolescentes | 0.39 | -4.10 | -1.70 | 1.92 | | | | |
| Adolescentes | 0.35 | -6.17 | -1.41 | 3.88 | | | | |
| Item 14 | | | | | 0.02822 | 0.01350 | 0.00559 | .001 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Preadolescentes | 1.08 | 0.01 | 1.45 | 2.19 | | | | |
| Adolescentes | 0.91 | -0.38 | 1.44 | 3.16 | | | | |
| Item 15 | | | | | -0.00648 | 0.00208 | 0.00577 | ns |
| Preadolescentes | 1.02 | -0.96 | -0.99 | 2.39 | | | | |
| Adolescentes | 1.22 | -0.85 | 0.84 | 2.18 | | | | |
| Item 16 | | | | | -0.01058 | 0.00773 | 0.00620 | .005 |
| Preadolescentes | 1.12 | -1.19 | 0.52 | 1.43 | | | | |
| Adolescentes | 0.86 | -1.60 | 0.95 | 2.32 | | | | |
| Item 17 | | | | | 0.09355 | 0.14911 | 0.01323 | .001 |
| Preadolescentes | 0.38 | -0.74 | 1.44 | 4.33 | | | | |
| Adolescentes | 0.21 | -3.68 | -0.21 | 5.21 | | | | |
| Item 18 | | | | | -0.01073 | 0.00241 | 0.00331 | ns |
| Preadolescentes | 1.41 | -0.83 | 1.05 | 1.96 | | | | |
| Adolescentes | 1.78 | -0.85 | 1.09 | 2.37 | | | | |
| Item 19 | | | | | -0.04527 | 0.03240 | 0.00511 | .001 |
| Preadolescentes | 1.16 | -1.79 | -0.41 | 1.60 | | | | |
| Adolescentes | 1.33 | -1.52 | -0.05 | 1.94 | | | | |
| Item 20 | | | | | 0.00807 | 0.00168 | 0.00335 | ns |
| Preadolescentes | 1.07 | -0.51 | 1.50 | 2.55 | | | | |
| Adolescentes | 1.10 | -0.82 | 1.68 | 3.08 | | | | |
| Item 21 | | | | | 0.04326 | 0.02886 | 0.00740 | .001 |
| Preadolescentes | 0.71 | -0.41 | 1.69 | 3.18 | | | | |
| Adolescentes | 0.66 | -1.13 | 1.33 | 3.50 | | | | |
| Item 22 | | | | | 0.00461 | 0.00194 | 0.00588 | ns |
| Preadolescentes | 0.98 | -0.32 | 0.73 | 2.52 | | | | |
| Adolescentes | 0.96 | -0.77 | 0.99 | 3.81 | | | | |
| Item 24 | | | | | -0.00657 | 0.00089 | 0.00452 | ns |
| Preadolescentes | 0.82 | -0.24 | 1.60 | 2.84 | | | | |
| Adolescentes | 0.88 | -0.52 | 1.78 | 3.98 | | | | |
| Item 25 | | | | | -0.00725 | 0.00127 | 0.00492 | ns |
| Preadolescentes | 0.67 | 1.29 | 3.03 | 3.99 | | | | |
| Adolescentes | 0.57 | 1.40 | 3.97 | 6.04 | | | | |
| Item 26 | | | | | 0.03640 | 0.01927 | 0.01010 | .005 |
| Preadolescentes | 0.25 | -0.92 | 6.26 | 10.31 | | | | |
| Adolescentes | 0.29 | -1.45 | 4.11 | 7.26 | | | | |
| Item 27 | | | | | -0.01764 | 0.00469 | 0.00653 | ns |
| Preadolescentes | 0.63 | -1.64 | 1.20 | 2.92 | | | | |
| Adolescentes | 0.66 | -1.51 | 1.31 | 3.38 | | | | |
| Item 29 | | | | | -0.04831 | 0.03303 | 0.00741 | .001 |
| Preadolescentes | 0.65 | -1.29 | 1.43 | 3.01 | | | | |
| Adolescentes | 0.45 | -0.83 | 2.92 | 5.63 | | | | |

*Punto de corte establecido del índice NCDIF con el procedimiento IPR para α =.01

Se ha obtenido un valor de DTF igual a 0.08226, lo que indica que la escala presenta equivalencia de medida entre ambos grupos de edad, ya que el criterio establecido para la

escala completa a partir del cual se considera que hay funcionamiento diferencial en este conjunto de datos es de 0.16619, muy por encima del valor de DTF hallado.

En la Figura 3.39 se muestran las CCT para ambos grupos de edad, no apreciándose diferencias en la puntuación esperada del test entre preadolescentes y adolescentes.

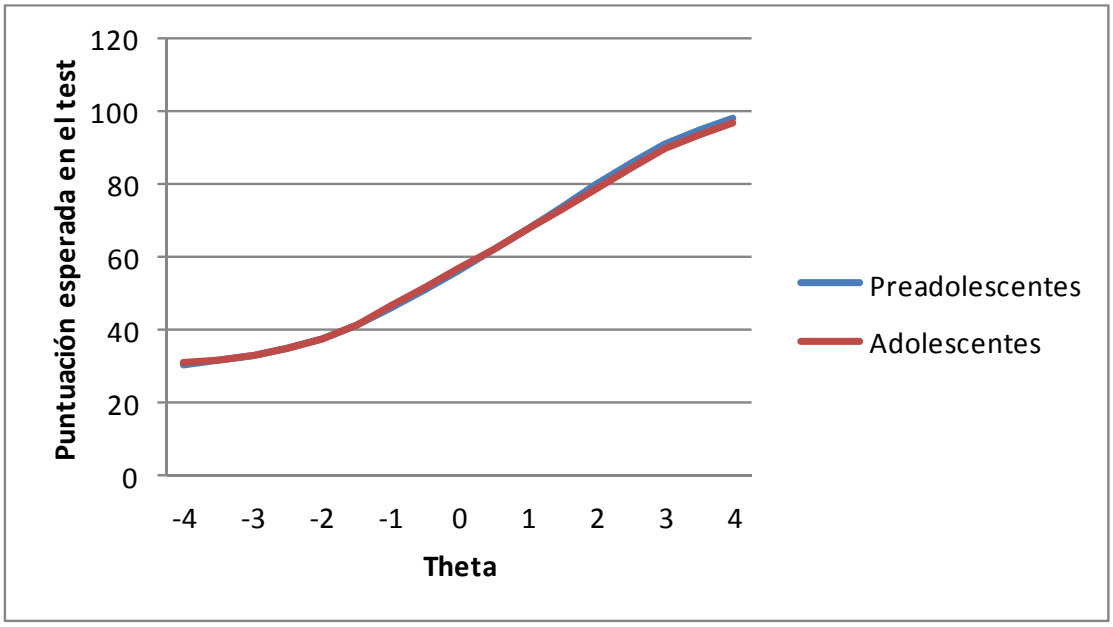



Figura 3.39. CCT para preadolescentes y adolescentes en el test BIS.

A pesar de que a nivel de la escala completa no hay indicios de funcionamiento diferencial, encontramos que más de la mitad de los ítems presentan funcionamiento diferencial, dado que su valor de NCDIF es significativo (α=.01). Así, de los 27 ítems que conforman la escala completa 16 presentan DIF: los ítems 1, 2, 3, 4, 6, 8, 9, 10, 11, 14, 16, 17, 19, 21, 26 y 29.

En los análisis por subescalas (ver apartados 3.5.2.1, 3.5.2.2. y 3.5.2.3) se encuentran resultados similares:

(1) En la subescala IM coinciden en detectar los ítems 2, 26 y 29, encontrado además dos ítems adicionales con DIF (el ítem 6 y el 9).

(2) En la subescala INP en ambos casos se detectan 6 ítems, pero hay uno que no es coincidente: el ítem 14 presenta DIF en el análisis de la escala completa pero no en el de las subescalas, y el ítem 22 es el caso contrario, ya que presenta DIF en el análisis de la subescala pero no en de la escala completa.

(3) Lo mismo sucede en la subescala ICA en la que coinciden en detectar los ítems 4, 10, 19 y 21, pero no en el ítem 16 detectado únicamente a nivel de test completo, ni en el ítem 13, detectado únicamente a nivel de subescala.

Los resultados que se han mostrado aquí distan mucho de los encontrados con la primera versión de este software (DFITP5), sobre todo en la detección del funcionamiento diferencial a nivel de ítem. El estudio de la equivalencia de medida entre hombres y mujeres indica que no hay funcionamiento diferencial en ninguna de las tres subescalas de Impulsividad. Además, siguiendo el punto de corte establecido por Flowers *et al.* (1999), ningún ítem presenta DIF, ya que ninguno tiene un valor de NCDIF igual o superior a 0.054. Los resultados de este procedimiento en relación a la variable edad tampoco muestran funcionamiento diferencial del test en ninguna de las tres subescalas, y sólo tres ítems tienen DIF: el ítem 4, el ítem 10 y el ítem 17.

Los resultados obtenidos confirman, como era de esperar, una mayor tasa de detección de funcionamiento diferencial de los ítems utilizando la versión DFIT8 del software (ver Tabla 3.82 para una comparación de ambas versiones).

Tabla 3.82. *Funcionamiento diferencial de ítems y tests de las tres subescalas del BIS y de la escala completa, relativos a las variables sexo y edad*

| | **SEXO** | | **EDAD** | |
|---|---|---|---|---|
| | **DFITP5** | **DFIT8** | **DFITP5** | **DFIT8** |
| Subescala IM | No DTF | No DTF | No DTF | No DTF |
| Ítem 2 | | | | DIF |
| Ítem 6 | | | | |
| Ítem 9 | | | | |
| Ítem 12 | | | | |
| Ítem 15 | | | | |
| Ítem 18 | | | | |
| Ítem 26 | | | | DIF |
| Ítem 29 | | DIF | | DIF |
| Subescala INP | No DTF | No DTF | No DTF | DTF (elimina ítem 7) |
| Ítem 1 | | | | DIF |
| Ítem 3 | | DIF | | DIF |
| Ítem 5 | | | | |
| Ítem 8 | | DIF | | DIF |
| Ítem 11 | | | | DIF |
| Ítem 14 | | | | |
| Ítem 17 | | | DIF | DIF |
| Ítem 22 | | DIF | | DIF |
| Ítem 25 | | DIF | | |
| Subescala ICA | No DTF | No DTF | No DTF | No DTF |
| Ítem 4 | | | DIF | DIF |
| Ítem 7 | | DIF | | |
| Ítem 10 | | DIF | DIF | DIF |
| Ítem 13 | | DIF | | |
| Ítem 16 | | | | |
| Ítem 19 | | | | DIF |
| Ítem 20 | | | | |
| Ítem 21 | | | | DIF |
| Ítem 24 | | DIF | | |
| Ítem 27 | | DIF | | |
| **TOTAL ESCALA** | **No DTF** | **No DTF** | **No DTF** | **No DTF** |

# Sección III. CONCLUSIÓN Y DISCUSIÓN

Los resultados obtenidos en el estudio permiten concluir que, al examinar la equivalencia de medida, los tres procedimientos evaluados arrojan resultados semejantes en cuanto a los ítems que presentan funcionamiento diferencial (en especial, los dos métodos basados en la TRI), resultados que difieren cuando se trata de evaluar el funcionamiento diferencial a nivel de escala. Por otro lado, mientras que en la variable sexo se puede defender la equivalencia parcial de medida entre hombres y mujeres, la mayoría de los procedimientos estadísticos abundan en la falta de equivalencia psicométrica entre los dos grupos de edad aquí considerados.

En el caso de la equivalencia de medida entre sexos, los distintos procedimientos aportan evidencia a favor de la existencia de algún grado de equivalencia de medida entre hombres y mujeres, aunque no coinciden por completo en su discriminación de los ítems que presentan DIF.

No hay equivalencia métrica total de medida al trabajar con el AFC, aunque únicamente es necesario eliminar las restricciones de igualdad de cargas factoriales del ítem 1 para considerar la equivalencia métrica parcial de medida. También hay evidencias de equivalencia escalar parcial entre hombres y mujeres, una vez eliminadas las restricciones de igualdad de interceptos de los ítems 1, 6, 8, 10, 13, 24, 25, 27 y 29 del test.

Los dos procedimientos basados en la TRI difieren en su análisis del funcionamiento diferencial del test. Según los resultados del procedimiento basado en comparación de modelos, tanto en su modalidad más restrictiva de igualdad de *a* y *b*, como en su modalidad menos restrictiva de igualdad de *a*, hay funcionamiento diferencial del

test; sin embargo utilizando el procedimiento DFIT, los resultados indican invarianza de medida del test BIS según sexo.

El análisis de funcionamiento diferencial por subescalas revela que, en la subescala Impulso Motor, los dos procedimientos de comparación de modelos menos exigentes -que fuerzan la igualdad únicamente de λ y de *a*- y el procedimiento DFIT coinciden en que hay equivalencia de medida entre hombres y mujeres, no llegando a la misma conclusión los dos procedimientos de comparación de modelos más restrictivos. En la subescala Impulso No Planificado, únicamente el procedimiento DFIT halla equivalencia total de medida en las subescalas, encontrando ambos procedimientos de comparación de modelos (en sus dos versiones) equivalencia parcial de medida. En la subescala Impulso Cognitivo-Atencional, tanto el procedimiento de comparación de modelos basado en el AFC, en su versión de igualdad de cargas factoriales, como el procedimiento DFIT apuntan a la equivalencia entre sexos.

En cuanto a la diferencia en la detección del funcionamiento diferencial de los ítems, los tres procedimientos detectan un número similar de ítems con DIF: 9 con el procedimiento basado en el AFC multigrupo y 7 y 8 ítems con el procedimiento DFIT y la comparación de los parámetros de la TRI, respectivamente.

Hay una mayor similitud entre los dos procedimientos de la TRI en la detección de ítems con DIF (comparado con el AFC), coincidiendo ambos en 5 de los ítems detectados (ítems 3, 8, 13, 25 y 29). Hay tres ítems (16, 17 y 26) que son detectados por el procedimiento de comparación de modelos, pero no por el procedimiento DFIT, y otros

dos ítems (11 y 27) que son detectados por el procedimiento DFIT, pero no por el de comparación de modelos.

En el caso de la invarianza de medida entre preadolescentes y adolescentes únicamente el procedimiento DFIT avala esta equivalencia (aunque no en todas las subescalas), mostrando el resto de los procedimientos resultados que difícilmente pueden interpretarse como compatibles con la equivalencia de medida.

Según el AFC, no hay equivalencia métrica de medida, siendo necesario eliminar las restricciones de tres ítems de la escala (29, 8 y 14) para considerar la equivalencia parcial de medida. En el caso del modelo más restrictivo de equivalencia escalar es necesario eliminar la restricción de 15 ítems para lograr la equivalencia parcial, lo que supone más de la mitad de la escala. Los procedimientos basados en la TRI difieren en su análisis del funcionamiento diferencial del test. Según los resultados del procedimiento de comparación de modelos, tanto en su modalidad más restrictiva de igualdad de *a* y *b*, como en su modalidad menos restrictiva de igualdad de *a*, hay funcionamiento diferencial del test y de las tres subescalas; sin embargo utilizando el procedimiento DFIT los resultados indican invarianza de medida a nivel de escala en el test completo y en todas las subescalas a excepción de la subescala Impulso No Planificado. Por tanto, ninguno de los tres procedimientos abordados establece la equivalencia de medida en la subescala de Impulso No Planificado.

Por subescalas, los tres procedimientos encuentran menos presencia de ítems con funcionamiento diferencial en la subescala IM, con 3 detecciones. En la subescala INP el número de ítems con DIF es el más elevado, detectando 6 casos el procedimiento basado

en AFC y el procedimiento DFIT y 8 casos el procedimiento de comparación de modelos basado en la TRI. En cuanto a la subescala ICA, se detectan 6, 7 y 5 casos respectivamente para los procedimientos basados en el AFC, en la comparación de modelos basado en la TRI y DFIT.

En el test completo, el procedimiento de AFC detecta 15 ítems con DIF frente a los 20 y 16 que arrojan respectivamente la comparación de modelos basada en TRI y el procedimiento DFIT. Estas cifras, en cualquier caso, son notablemente superiores a las de la variable sexo, ya que suponen un porcentaje de ítems con DIF de más de la mitad de los ítems en todos los procedimientos utilizados.

Los resultados de las tres técnicas reflejan algunas similitudes en la detección de DIF. Hay 13 ítems que presentan DIF según ambos procedimientos basados en la TRI (los ítems 1, 2, 3, 4, 8, 10, 11, 14, 16, 17, 19, 21 y 27), siete que son detectados por el procedimiento de comparación de modelos pero no por el procedimiento DFIT (los ítems, 5, 13, 18, 20, 22, 24 y 25) y tres detectados por el procedimiento DFIT pero no por el de comparación de modelos (ítems 6, 9 y 27).

De manera global, estos resultados reflejan que todos los procedimientos excepto DFIT son muy exigentes a nivel de escala, descartando la equivalencia total de medida en la gran mayoría de los casos, obteniéndose únicamente una equivalencia parcial. Estos resultados concuerdan con los encontrados por Meade y Lautenschlager (2004c) en la detección de funcionamiento diferencial a nivel de escala.

Los resultados encontrados aquí no confirman la similitud esperable entre los procedimientos de comparación de modelos en estudios de equivalencia, dada la similitud que existe entre el parámetro de discriminación de la TRI y la carga factorial del AFC. Posiblemente esta incongruencia esté relacionada con el hecho de que los resultados no son directamente comparables, ya que en el procedimiento de comparación de modelos basado en el AFC se tuvo en cuenta la estructura trifactorial de la escala y en el basado en la TRI no. En relación a la variable sexo, el ítem que presenta DIF utilizando el procedimiento de comparación de modelos basado en el análisis factorial confirmatorio no es detectado también por el procedimiento de comparación de modelos (basado en la igualdad de *a*). En la variable edad hay 3 ítems que presentan DIF en el modelo basado en el AFC, y solo uno de ellos coincide con los ítems obtenidos del modelo equivalente en la TRI.

Resultados similares encuentran Meade y Lautenschlager (2004a), que hipotetizan un alto acuerdo en relación a la detección del DIF no uniforme en estas dos mismas técnicas mediante un estudio de simulación, y no pueden confirmarlo, concluyendo que el procedimiento basado en el AFC no detecta de manera adecuada ni los ítems con diferencias únicamente en el parámetro *b* ni los que difieren en el parámetro *a*.

Utilizando datos reales, Kim *et al*., (2010) encontraron, en el ámbito de la equivalencia parcial de medida, que el AFC detectó más casos de ítems con DIF que el procedimiento de comparación de modelos de la TRI, lo que concuerda con las hipótesis de Raju *et al.* (2002). Nuestros resultados no apoyan este punto ya que, al igual que los resultados de otras investigaciones (por ej. Reise *et al*., 1993), encontramos un número similar de ítems con DIF, aunque algo mayor utilizando algún procedimiento basado en la TRI. En esta línea están los resultados de Scandura *et al.* (2001), que encuentran

equivalencia parcial de medida utilizando el AFC multigrupo pero no bajo la aproximación de comparación de modelos de la TRI.

Otro foco de discrepancias entre los diversos procedimientos es el grado de acuerdo en la detección de ítems invariantes entre los procedimientos. Algunos estudios encuentran grandes discrepancias entre las detecciones de las distintas técnicas (Kim, *et al*., 2010; Reise *et al*., 1993; Scandura *et al*., 2001), mientras que nuestros resultados encuentran un grado aceptable de acuerdo entre las tres técnicas empleadas, resultados también respaldados por algunos estudios en la literatura (Facteau y Craig, 2001; Maurer *et al.*, 1998; Raju *et al*., 2002).

El por qué de estas discrepancias es una cuestión clave que todavía no está resuelta en la literatura. Ahondando en los estudios citados que apoyan la convergencia entre los procedimientos para buscar una explicación a sus resultados, se observa que, por ejemplo, el estudio de Facteau y Craig (2001) no fue demasiado exigente a la hora de establecer la equivalencia de medida con el AFC multigrupo y utilizó la versión DFIT5P del software en el procedimiento DFIT, que es poco sensible en la detección de casos con DIF. La práctica totalidad de los estudios revisados utilizan los puntos de corte para NCDIF y DTF establecidos por Raju, van der Linden y Fleer en 1995, por lo que probablemente con la última versión del programa (DFIT8) se encontrarían un mayor número de ítems con DIF. DFITP8 utiliza un test de significación denominado método de replicación parámetro ítem (IPR) desarrollado recientemente por Oshima *et al*. (2006) que es más sensible en la detección de ítems con DIF.

En el presente estudio se ha podido comprobar que utilizando la versión DFITP5 del software los resultados distan mucho de los encontrados con el resto de procedimientos para evaluar la equivalencia de medida entre grupos, siendo éste el menos sensible para detectar DIF y DTF de los tres procedimientos utilizados. En la literatura especializada que compara diversos procedimientos de equivalencia de medida (ver por ejemplo Meade y Lautenschlager, 2004; Monnot y Griffith, 2005), también se concluye que la versión DFITP5 del programa detecta menos casos positivos tanto a nivel de ítem como a nivel de escala.

Otro factor a tener en cuenta a la hora de explicar las discrepancias en los resultados es la sensibilidad al tamaño muestral de los distintos métodos utilizados. El procedimiento de comparación de modelos basado en la TRI se basa en el test de razón de verosimilitud, que está afectado por el tamaño muestral, razón por la cual este procedimiento resulta muy exigente en muestras grandes que, por otra parte, son necesarias para abordar la estimación de parámetros de la TRI. Los estudios de comparación de modelos basados en el AFC también suelen utilizar un estadístico que está muy influido por el tamaño muestral ($\Delta\chi^2$), lo que puede paliarse complementando su uso con el $\Delta$CFI. Esta medida, tomada en esta investigación, todavía no se ha adoptado de manera generalizada en los estudios revisados.

Resulta coherente, por otra parte, que el AFC difiera algo más en sus resultados, dado que se ha desarrollado en términos de un modelo multidimensional. El instrumento de medida utilizado en este trabajo tiene una estructura trifactorial, aunque su ajuste al modelo unidimensional es más que razonable. En el procedimiento basado en el AFC se tiene en cuenta la estructura tridimensional de la escala en los estudios de equivalencia entre

grupos, mientras que los procedimientos basados en la TRI tratan por separado cada subescala y posteriormente el test completo de manera unidimensional. Esto hace que los resultados TRI sean perfectamente comparables, pero que haya que ser cauteloso al compararlos con los resultados del AFC. Sería deseable en posteriores investigaciones la utilización de modelos multidimensionales de la TRI para el examen de la equivalencia de medida con su homólogo basado en el AFC.

El AFC difiere también de los métodos basados en la TRI en el número de parámetros estimados, mayor en el segundo caso. En el AFC multigrupo se estiman, para cada ítem, la carga factorial, el intercepto y el término unicidad mientras que en los procedimientos basados en la TRI se estima para cada ítem un parámetro *a* de discriminación y tantos parámetros *b* como el número de alternativas menos uno. La información adicional que aportan los parámetros *b* de los procedimientos basados en la TRI plantean condiciones de equivalencia más exigentes.

Por otro lado, hay también algunas diferencias relevantes entre los dos procedimientos basados en la TRI, al utilizar la información obtenida por esos parámetros adicionales de manera diferente. El test de razón de verosimilitud compara los parámetros de cada ítem con todas las condiciones posibles de los datos para establecer la equivalencia. Este examen es el más riguroso, ya que indica presencia de DIF si cualquiera de los parámetros difiere en un único ítem. Por el contrario, el procedimiento DFIT utiliza una aproximación más pragmática en la evaluación del DIF: si los parámetros de un ítem varían solo en las personas con un nivel muy alto o muy bajo del rasgo latente y hay pocas respuestas observadas para esas opciones extremas de respuesta, estas diferencias tendrán un impacto mínimo en el cómputo global. Como resultado de esta propiedad del

procedimiento DFIT, y de la utilización de un punto de corte para valorar el funcionamiento diferencial en lugar de tests paramétricos más estrictos, el procedimiento DFIT puede resultar menos exigente que el test de razón de verosimilitud en la detección de DIF, pero también más realista en cuanto a la importancia en las detecciones de funcionamiento diferencial encontradas.

La principal ventaja del procedimiento DFIT es la compensación a nivel de test de los ítems que presentan DIF en distinta dirección. Es el único de los procedimientos que abordan la equivalencia de medida cuyo índice global, DTF, se basa en el funcionamiento diferencial compensatorio de los ítems de la escala -el índice CDIF-, de manera que si un ítem influye a favor del grupo 1 y otro ítem influye de igual forma, pero a favor del grupo 2, el CDIF sumado de estos dos ítems se compensará cuando se combinen para formar el DTF del test total. Esto hace que sea un método más conveniente cuando se vayan a utilizan las puntuaciones totales de una escala y subescala de modo convencional, en contraposición a ítems procedentes de un banco en la aplicación de un Test Adaptativo Informatizado (TAI).

Varios autores consideran que, en general, funcionan mejor los procedimientos basados en la TRI, en el sentido de que proporcionan información más útil para establecer la equivalencia de medida (Breithaupt y Zumbo, 2002; Flowers *et al*., 2002; Maurer *et al.*, 1998; Meade y Lautenschlager, 2004a; McDonald, 1999; Raju *et al.*, 2002), al detectar si la fuente del DIF se atribuye a diferencias en el parámetro *a* o en el parámetro *b*. De esta manera, el análisis de equivalencia de medida teniendo en cuenta estos dos parámetros proporciona más información sobre el funcionamiento de los ítems y permite establecer

conclusiones más precisas sobre la equivalencia de medida de la prueba que la aproximación con análisis factorial confirmatorio.

En cualquier caso, teniendo en cuenta que los procedimientos utilizados para valorar la equivalencia psicométrica proporcionan información diferente, puede resultar recomendable utilizar, de manera complementaria, distintas técnicas basadas en el AFC y en la TRI. En este sentido, sus resultados podrían considerarse en una interpretación eminentemente práctica como piezas de información sobre la equivalencia de medida de una prueba, en la línea de acumular evidencias de validez. De este modo, con cada prueba de invarianza que apoye la equivalencia entre los grupos, los investigadores y los profesionales pueden estar más seguros de que sus pruebas están funcionando de manera equivalente. Si el procedimiento DFIT, el test de razón de verosimilitud y las pruebas de AFC indican invarianza entre los grupos, se pueden realizar comparaciones entre los grupos con un alto grado de certeza. Si solo uno de los procedimientos indican equivalencia de medida, las conclusiones de las comparaciones entre los grupos deben hacerse con cautela, siempre teniendo en cuenta que a nivel de test los dos procedimientos de comparación de modelos son mucho más exigentes que el procedimiento DFIT.

Aunque no son numerosos, hay varios ejemplos de estudios que utilizan ambas metodologías de manera complementaria (Schmit, Kihm y Robie, 2000; Facteau y Craig, 2001; Maurer *et al.*, 1998; Scandura *et al*., 2001; Zickar y Robie, 1999). Incluso hay autores que utilizan el procedimiento DFIT como medida global del funcionamiento diferencial del test, recurriendo a otros procedimientos para evaluar el funcionamiento diferencial de cada ítem individual, (Cooke *et al*., 2001). Sin embargo, realizar los dos tipos de análisis puede no ser posible en todos los casos debido a cuestiones prácticas o de

otra índole. En estos casos es fundamental tener en cuenta factores como el tamaño muestral, la dimensionalidad del test, los recursos disponibles y la utilización que se va a hacer de las puntuaciones de los tests empleados, para tomar una decisión sobre la metodología más apropiada en cada caso.

Respecto al tamaño muestral, dado que el número de parámetros a estimar es menor en el AFC, este procedimiento resulta preferible para analizar la equivalencia cuando el tamaño muestral es bajo.

Respecto a la dimensionalidad de la prueba, el análisis factorial confirmatorio proporciona información sobre la relación entre los factores latentes, por lo que su utilización será imprescindible cuando el objetivo de la investigación sea examinar la equivalencia en una prueba multifactorial. Si el interés recae en evaluar la equivalencia en un instrumento de medida unidimensional o en un conjunto de ítems de una escala puede resultar más apropiado utilizar un procedimiento basado en la TRI, ya que proporciona más información a nivel de ítem, y no hay relaciones entre los factores que analizar.

En cuanto a los recursos disponibles es más sencillo llevar a cabo un análisis factorial confirmatorio cuando hay que comparar muchos grupos que utilizar un procedimiento basado en la TRI, ya que en el análisis factorial confirmatorio se comparan todos los grupos simultáneamente mientras que el análisis basado en la teoría de respuesta al ítem requiere comparaciones por pares.

Respecto a la utilización que se va a hacer de las puntuaciones obtenidas por el test hay que tener en cuenta la distinta sensibilidad de los procedimientos utilizados para

encontrar DIF. En este sentido, hay que valorar la importancia que pueden tener las detecciones erróneas según las consecuencias que tengan para los sujetos la utilización prevista de las puntuaciones del test. No cabe duda de la importancia que está adquiriendo el estudio de las consecuencias, ya que en la última edición disponible de los *Standards for Educational and Psychological Testing* (AERA, APA y NCME, 1999), y en la próxima que está en fase de publicación, se han incluido la validación de las consecuencias del uso de los tests como una fuente más de evidencia de validez.

El test BIS es uno de los instrumentos de medida de la impulsividad más utilizado (Arce y Santisteban, 2006). Se ha administrado para obtener información acerca de la prevalencia de esta característica en una determinada población, y también con una función diagnóstica en estudios de desórdenes bipolares (e.g. Swann, Pazzaglia, Nicholls, Dougherty y Moeller, 2003), alcohol y abuso de sustancias (e.g., Moeller, Dougherty, Barratt, Schmitz, Swann y Grabowski, 2001), y desórdenes de personalidad (e.g., Soloff, Kelly, Strotmeyer, Malone y Mann, 2003), entre otros. Cuando la prueba se utiliza con fines diagnósticos, es preferible utilizar una técnica muy sensible a la detección de DIF para evitar que variables como el sexo y la edad del sujeto puedan interferir en un correcto diagnóstico del sujeto. Obviamente, si la función del test es descriptiva no se derivarán consecuencias sobre los sujetos a los que se le aplica, por lo que éste no será un factor relevante para decidir qué procedimiento de equivalencia utilizar.

Otra cuestión relevante consiste en considerar hasta qué punto los resultados encontrados en cuanto al funcionamiento diferencial de la escala tienen realmente una importancia práctica a la hora de utilizar la escala. En muestras muy amplias, como la de esta investigación, se puede encontrar funcionamiento diferencial estadísticamente

significativo aunque los efectos de significación práctica sean pequeños (Kirk, 1996). En el presente estudio se ha utilizado como estrategia para calibrar la importancia práctica la representación de las curvas características del test en cada grupo. Así, se constata que en la mayoría de los casos en los que los dos procedimientos basados en comparación de modelos (AFC y TRI) encuentran funcionamiento diferencial a nivel de escala, en los gráficos no se aprecian apenas diferencias en las puntuaciones esperadas en el test. Por tanto, resulta recomendable complementar las técnicas estadísticas con técnicas gráficas que aporten información visual clara sobre la importancia práctica del funcionamiento diferencial encontrado.

Un tema que se ha abordado muy poco, aún siendo una cuestión de indudable interés, es qué hacer con los ítems que presentan funcionamiento diferencial. Habitualmente, en los trabajos sobre DIF estos ítems son eliminados (Robie *et al*., 2001); sin embargo, dado que el enfoque de nuestro trabajo se dirige a la totalidad de la escala, y en ésta se ha comprobado la equivalencia parcial entre grupos, no está tan clara esta cuestión. Las implicaciones de la invarianza parcial para la interpretación de la medida han sido ampliamente ignoradas en la literatura científica (Millsap y Kwok, 2004).

Una opción sería la utilización de una versión reducida de la escala, que omita los ítems que funcionan de manera diferencial entre las dos poblaciones. Según Cheung y Rensvold (1998), esta opción presenta un importante inconveniente y es que en varios estudios de invarianza podrían crearse muchas versiones diferentes de una escala para distintas poblaciones. Roznowski (1987) también argumenta en contra de eliminar "a ciegas" los ítems que presentan funcionamiento diferencial en distintos grupos. Según su opinión, "la purificación del test por la eliminación de los ítems que presentan este

problema puede contribuir solo a que la homogeneidad del conjunto de ítems sea mayor, lo que podría disminuir la validez y la precisión predictiva y, paradójicamente, incrementar la contribución a la varianza total de otros determinantes no relacionados con el rasgo medido" (p.463). En esta línea, Drasgow y Hulin (1990) afirman que de los ítems que presentan DIF en una escala solo deben eliminarse los que contribuyen al funcionamiento diferencial del test, porque eliminar todos los ítems con DIF puede producir un grado de homogeneidad en la escala que degrada la validez predictiva.

Una segunda opción sería utilizar la escala completa, al considerar que las diferencias encontradas entre los grupos en la estructura factorial son pequeñas en cualquier sentido y que no perjudicarán las inferencias realizadas con la escala.

Por último, también podría abandonarse la utilización de esa misma escala en las comparaciones entre distintos grupos, basándonos en que la pérdida de equivalencia establece que la escala mide diferentes variables latentes en ambos grupos.

La literatura actual sobre invarianza factorial apenas ofrece orientación para elegir entre estas tres opciones. Zieky (1993) considera que la imparcialidad de un ítem se relaciona estrechamente con el propósito para el que se utiliza el test, siendo necesario incluir, en los estudios de DIF, juicios de expertos en desarrollo de tests y especialistas en la materia. Aconseja combinar el análisis teórico del ítem con el estadístico. Así, cada ítem de cada test desarrollado se sometería a un escrutinio por parte de revisores y especialistas entrenados que, siguiendo un amplio conjunto de directrices, se aseguraría de que los ítems no resultan ofensivos, no refuerzan estereotipos negativos y que las cuestiones son apropiadas para una sociedad multicultural. El análisis DIF no constituye en absoluto una

sustitución de este estudio teórico, es más, deberían eliminarse los ítems que no superen el estudio teórico aunque no muestren DIF. Según Zieky (1993), el análisis DIF constituye un seguro adicional para ayudarnos a garantizar la imparcialidad de los ítems del test.

Otros autores (e.g., Linn, 1993; Penfield y Lam, 2000; Rousos y Stout, 1996) enfatizan la necesidad de distinguir entre DIF y sesgo en esta cuestión. Así, se podría diferenciar entre DIF estadístico y DIF sustantivo. El DIF estadístico se refiere a la identificación estadística del DIF tal y como se haya definido, y el DIF sustantivo se refiere a la identificación del constructo que está dando lugar al DIF (es el responsable de las diferencias entre los grupos, sin que el ítem se haya diseñado para medirlo). Solo cuando un ítem presenta los dos tipos de DIF se considera sesgado y un firme candidato para su eliminación del test. Por tanto, los estudios de DIF no se deben utilizar para eliminar ítems directamente, siendo necesario tener en cuenta una valoración teórica del ítem.

La cuestión principal para tomar una decisión es la importancia de que cualquier violación de la invarianza factorial de una medida sea juzgada en relación al propósito de la medida. Esto es, estaría íntimamente ligado al concepto de validez. La pregunta crucial sería, ¿cuál es la utilización que se va a hacer de la escala en la práctica? Una vez se describa su uso entonces es cuando hay que preguntarse si las violaciones particulares de la invarianza interfieren con él. Esta cuestión está estrechamente relacionada con la justicia o equidad de las pruebas utilizadas. En este sentido, y con una orientación eminentemente práctica, Zieky (2006) proporciona un conjunto de 16 directrices a seguir para tratar de asegurar que los tests construidos con fines de certificación o acreditación son justos.

Una de las limitaciones de este estudio tienen que ver con el escaso número de ítems que componen las subescalas del test BIS, ya que todos los procedimientos abordados -y en especial los basados en la TRI- necesitan de una cantidad moderada de ítems que proporcionen estimaciones adecuadas del nivel de rasgo del sujeto.

Pese a contar con las ventajas de utilizar una muestra probabilística y representativa de la población en una aplicación real del test, resulta una desventaja evidente no saber cuáles son los ítems que inequívocamente presentan DIF, por lo que sería necesario realizar un estudio de simulación que permita determinar la eficacia real de cada una de las metodologías abordadas en este trabajo.

Los trabajos de simulación consultados difieren en los procedimientos utilizados, lo que hace complicada su comparación. Además están mayoritariamente dedicados al análisis del funcionamiento diferencial a nivel de ítem, habiendo pocos estudios centrados en el estudio del test completo. En cuanto a las variables que contemplan, las más habituales son tipo y cantidad de DIF, tamaño muestral, número de categorías de respuesta y cantidad de impacto. Como continuación del presente trabajo se propone un estudio de simulación cuyo objetivo fundamental sea comprobar la eficacia de los tres procedimientos aquí abordados para evaluar la equivalencia de medida de un test, incluyendo como variable relevante del estudio el número de ítems de la prueba.

Una línea de investigación muy interesante que trata de hacer más comparables los procedimientos basados en AFC y TRI es la iniciada recientemente por Kim y Yoon (2011), que utilizan un AFC multigrupo para categorías ordenadas con una estructura de umbrales, similar a la que utiliza el modelo de respuesta graduada de Samejima en la TRI.

No cabe duda que es necesario seguir investigando sobre la equivalencia de medida en varios grupos mediante distintos procedimientos (Meade y Lautenschlager, 2004a; Raju *et al*., 2002; Reise, Widaman y Pugh, 1993). Según Vandenberg (2002) no solo es necesario continuar investigando sobre invarianza factorial desde el punto de vista de los procedimientos analíticos subyacentes, sino también sobre su aplicabilidad. En otras palabras, hay que ahondar en nuestro conocimiento sobre las condiciones que hacen la invarianza más apropiada y las consecuencias de sus limitaciones con respecto a su aplicación. Riordan, Richardson, Schaffer y Vandenberg (2001) consideran necesario un incremento de las investigaciones sobre el tema, en especial de los estudios de Monte Carlo para determinar la eficacia de la metodología existente para investigar la invarianza. Tal y como ha comentado Little (2000), hay muchas cuestiones que han de ser consideradas y examinadas antes de utilizar estos procedimientos de manera inequívoca.

Esta es, en realidad, la conclusión a la que apuntan los resultados de la presente investigación: una razonable consistencia en la detección a nivel de ítem que no se obtiene cuando se trabaja a nivel de test y que hace plantearse la conveniencia de complementar con un análisis gráfico los resultados procedentes del análisis estadístico, además de utilizar más de un procedimiento para analizar la equivalencia.

En cualquier caso, conviene tener siempre presente que se está trabajando con modelos, y "ningún modelo es totalmente fiel a la conducta bajo estudio. Los modelos son, por lo general, formalizaciones de procesos que son extremadamente complejos. Es un error hacer caso omiso de cualquiera de sus limitaciones o de su artificialidad. Lo mejor que se puede esperar es que algún aspecto de un modelo pueda ser útil para la descripción, la predicción o la síntesis" (Cudeck y Henly, 1991, p. 521).

# Referencias

Abad, F. J., Olea, J., Ponsoda, V., y García, C. (2011). *Medición en ciencias sociales y de la salud*. Madrid: Síntesis.

Ackerman, T. (1992). A Didactic Explanation of Item Bias, Item Impact, and Item Validity from a Multidimensional Perspective. *Journal of Educational Measurement*, *29*(1), 67-91.

Alvarado, J. M., y Santisteban, C. (2006). *La validez en la medición psicológica*. Madrid: UNED.

American *Educational* Research Association, American *Psychological* Association, y National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American *Educational* Research Association, American *Psychological* Association, y National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American *Educational* Research Association.

American *Psychological* Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin Supplement, 51*(2), 1-38.

American *Psychological* Association, American *Educational* Research Association, y National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American *Psychological* Association.

American *Psychological* Association, American *Educational* Research Association, y National Council on Measurement in Education. (1974). *Standards for educational and psychological tests*. Washington, DC: American *Psychological* Association.

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1-15.

Angoff, W. (1988). Validity: An evolving concept. *Test validity* (pp. 19-32). Hillsdale, NJ England: Lawrence Erlbaum Associates, Inc.

Arce, E., y Santisteban, C. (2006). Impulsividad: Una revisión. *Psicothema*, *18*(2), 213-220.

Atkinson, L. (1988). The measurement - statistics controversy: Factor analysis and subinterval data. *Bulletin of the Psychonomic Society, 26*, 361-364.

Ávalo, J., Lévy, J. P., Rial, A., y Varela, J. (2006). Invarianza factorial con muestras múltiples. En J. P. Lévy, y J. Varela (Eds.), *Modelización con estructuras de covarianzas en ciencias sociales. Temas esenciales, avanzados y aportaciones especiales*. A Coruña: Netbiblo.

Babakus, E., Ferguson, C. E., y Joreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research, 24*, 222-228

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16* (1), 87-96.

Baker, F. B. (1995). *EQUATE 2.1: Computer program for equating two metrics in item response theory*. Madison: University of Wisconsin, Laboratory of Experimental

Design.

Barbero, I., Vila, E., y Holgado, F. P. (2010). *Psicometría*. Madrid: Sanz y Torres.

Barratt, E. S. (1959). Anxiety and impulsiveness related to psychomotor efficiency. *Perceptual and Motor Skills*, *9*, 191-198.

Barratt, E. S. (1994). Impulsiveness and aggression. En J. Monahan y H. J. Steadman (Eds.), *Violence and mental disorders, developments in risk assessment*. Chicago: The University of Chicago Press.

Batista, J. M. y Coenders, G. (2000). *Modelos de Ecuaciones Estructurales*. Madrid: La Muralla.

Batista-Foguet, J.M., Coenders, G. y Alonso, J. (2004). Análisis factorial confirmatorio. Su utilidad en la validación de cuestionarios relacionados con la salud. *Medicina Clínica, 122 (Suplemento)*, 21-27.

Baylé, F. J., Bourdel, M. C., Caci, H., Gorwood, P., Chignon, J., Adés, J., y Lôo, H. (2000). Structure factorielle de la traduction française de l'échelle d'impulsivité de barratt (BIS-10). *The Canadian Journal of Psychiatry / La Revue Canadienne De Psychiatrie,45*(2), 156-165.

Benson, J. (1987). Detecting item bias in affective scales. *Educational and Psychological Measurement, 47*(1), 55-67.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modelling. *Annual Review of Psychology, 31*, 419-456.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Bentler, P. M. (1992). On the fit of models to covariances and methodology to the Bulletin. *Psychological Bulletin, 112*, 400-404.

Bentler, P. M. (1995). *EQS Structural equations program manual*. Encino: Multivariate Software.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74*(1), 137-143.

Bentler, P. M. y Bonett, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin, 88* (3), 588-606.

Bentler, P. M. y Chou, C. P. (1987). Practical issues in structural modeling. *Sociological Methods & Research, 16*, 78-117.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee´s ability. En F. M. Lord y M. R. Novick (Eds). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bolt, D. M. (2002). A monte carlo comparison of parametric and nonparametric plytomous DIF detection methods. *Applied Measurement in Education, 15*(2), 113-141.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53* (1), 605-634.

Bollen, K. A. y Long, J. S., eds. (1993). *Testing structural equation models*. Newbury Park, CA: Sage.

Bolt, D. M., Hare, R. D., Vitale, J. E., y Newman, J. P. (2004). A Multigroup Item Response Theory Analysis of the Psychopathy Checklist-Revised. *Psychological Assessment, 16* (2), 155-168.

Borges, N., van den Bergh, B., y Hox, J. (2001). Testing measurement and structural equivalence in different age groups of children. *Kwantitatieve Methoden, 67*, 65-80.

Braddy, P. W., Meade, A. W., y Johnson, E. C. (2006). *Practical implications of using different tests o measurement invariance for polytomous measures*. Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Breckler, S. J. (1990). Applications of covariance structure modeling in Psychology: Cause for concern? *Psychological Bulletin, 107*, 260-271.

Brown, T. (2006). *Confirmatory factor analysis for applied research*. New York, NY US: Guilford Press.

Browne, M. W. y Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24* (4), 445-455.

Browne, M. W. y Cudeck, R. (1993). Alternative ways of assessing model fit. En K. A. Bollen y J. S. Long (Eds.), *Testing structural equation models*. Newbury Park, CA: Sage.

Budgell, G. R., Raju, N. S., y Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement, 19*(4), 309-321.

Byrne, B. M. (1994). Testing for the factorial validity, replication, and invariance of a measurement instrument: a paradigmatic application based on the Maslach Burnout inventory. *Multivariate Behavioral Research, 29* (3), 289-311.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: basic concepts, applications and programing*. New Jersey: Lawrence Erlbaum Associates.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema, 20*(4), 872-882.

Byrne, B. M., Shavelson, R. J., y Muthén, B. O. (1989). Testing for equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456-466.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues?. *Differential item functioning* (pp. 397-413). Hillsdale, NJ England: Lawrence Erlbaum Associates, Inc.

Camilli, G. y Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Candell, G. L. y Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12* (3), 253-260.

Cattell, J. K. (1893). Mental measurement. *Philosophical Review, 2*, 316-332.

Catell, R. B. (1981). *Personality and Learning Theory*. New York: Springer.

Chahin, N., Cosi, S., Lorenzo-Seva, U., y Vigil-Colet, A. (2010). Stability of the factor structure of Barrat's Impulsivity Scales for children across cultures: A comparison

of spain and colombia. *Psicothema, 22*(4), 983-989

Chan, D. (2000). Detection of differential item functioning on the Kirton Adaption-Innovation Inventory using multi-group mean and covariance structure analyses. *Multivariate Behavioral Research, 35*, 169-199.

Chang, H. H. y Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika, 59,* 391-404.

Chang, H., Mazzeo, J., y Roussos, L. A. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*(3), 333-353.

Chapple, C., y Johnson, K. (2007). Gender differences in impulsivity. *Youth Violence and Juvenile Justice*, *5*(3), 221-234.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464-504.

Chernyshenko, O. S., Stark, S., Chan, K., Drasgow, F., y Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*(4), 523-562.

Cheung, G. W., y Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*(1), 1-27.

Cheung, G. W. y Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9* (2), 233-255.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods.

*Multivariate Behavioral Research, 18*, 115-126.

Coenders, G., Batista Foguet, J. M., y Saris, W. E. (2005). *Temas avanzados en modelos de ecuaciones estructurales*. Madrid: La Muralla.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2ª ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cohen, A. S., Kim, S. H., y Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17* (4), 335-350.

Cohen, A. S., Kim, S. H., y Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20,* 15-26.

Cole, N. S. (1981). Bias in testing. *American Psychologist, 36* (10), 1067-1077.

Collins, W. C., Raju, N. S., y Edwards, J. E. (2000). Assessing differential functioning in a satisfaction scale. *Jounal of Applied Psychology, 85*(3), 451-461.

Conroy, D. E. y Motl, R. W. (2003). Modification, cross-validation, invariance, and latent mean structure of the self-presentation in Exercise Questionnaire. *Measurement in Physical Education and Exercise Science, 7* (1), 1-18.

Cooke, D. J., Kosson, D. S., y Michie, C. (2001). Psychopathy and ethnicity: Structural, item, and test generalizability of the Psychopathy Checklist--Revised (PCL-R) in Caucasian and African American participants. *Psychological Assessment, 13* (4), 531-542.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104.

Crocker, L. (2006). Introduction to measurement theory. In J. L. Green, G. Camilli y P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 371-384). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cronbach, L. J., y Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Cudeck, R. y Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18*, 147-167.

Cudeck, R., y Henly, S. J. (1991). Model selection in covariance structures analysis and the 'problem' of sample size: A clarification. *Psychological Bulletin, 109*(3), 512-519.

De Ayala, R.J. (2009). The Theory and Practice of Item Response Theory. NY: The Guildford Press.

DeShon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science, 46*(1), 137-149

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomous scored reading items under the generalized partial-credit model. *Journal of Educational Measurement, 31* (4), 295-311.

Dorans, N. J., y Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach*. Princeton, NJ: Educational,Testing Service.

Drasgow, F. (1982). Biased test items and differential validity. *Psychological Bulletin, 92,* 526-531.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72*(1), 19-29.

Drasgow, F. (1995a). Introduction to the polytomous IRT special issue. *Applied Psychological Measurement, 19*(1).

Drasgow, F. (1995b). Some comments on Labouvie and Ruetsch. *Multivariate Behavioral Research, 30* (*1*), 83-85.

Drasgow, F. y Hulin, C. L. (1990). Item response theory. En M. D. Dunnette y L. M. Hough (Eds.), *Handbook of industrial/organization psychology (2 ed)*. (pp. 577-636). Palo Alto, C.A.: Consultin Psychologists Press.

Drasgow, F. y Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populatins. *Journal of Applied Psychology, 70,* 662-680.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., y Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19* (*2*), 143-165.

Drasgow, F. y Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189-199.

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., y Tyler, R. (1951). *Intelligence and cultural differences; a study of cultural learning and problem-solving*. Chicago, IL US: University of Chicago Press.

Elosua, P. (2003). Sobre la validez de los tests. *Psicothema, 15*(2), 315-321.

Elosua, P. (2005). Evaluación progresiva de la invarianza factorial entre las versiones original y adaptada de una escala de autoconcepto. *Psicothema, 17*(2), 356-362.

Elosua, P. (2011). Measurement equivalence in ordered-categorical data. *Psicológica*, 32, 403-421.

Elosua, P., y Zumbo, B. D. (2008). Coeficientes de fiabilidad para escalas de respuesta categórica ordenada. *Psicothema, 20*(4), 896-901.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179-197.

Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist, 61*(1), 50-55.

Embretson, S. E. y Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Facteau, J. D. y Craig, S. B. (2001). Are performance appraisal ratings from different rating sources comparable? *Journal of Applied Psychology,* 86, 215-227.

Ferne, T., y Rupp, A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, *4*(2), 113-148.

Ferrando, P. J. (1996a). Calibration of invariant item parameters in a continuous item response model using the extended LISREL measurement model. *Multivariate Behavioral Research, 31*(4), 419-439.

Ferrando, P. J. (1996b). Evaluación de la unidimensionalidad de los ítems mediante análisis factorial. *Psicothema, 8*(2), 397-410.

Ferrando, P. J., y Lorenzo-Seva, U. (2005). IRT-related factor analytic procedures for testing the equivalence of paper-and-pencil and internet-administered

questionnaires. *Psychological Methods, 10*(2), 193-205.

Fidalgo, A. M. (1996). Funcionamiento diferencial de los items. En J. Muñiz (Coord.), *Psicometría*. Madrid: Universitas.

Flora, D. B., y Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466-491.

Flowers, C. P., Oshima, T. C., y Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement, 23* 309-326.

Flowers, C. P., Raju, N. S., y Oshima, T. C. (2002). *A comparison of measurement equivalence methods based on confirmatory factor analysis and item response theory.* Paper presented at NCME Annual Meeting.New Orleans.

Floyd, F. L. y Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, 7, 286-299.

Fontaine, J. R. (2005). Equivalence. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 803-813). San Diego: Academic Press.

Fossati, A., Barratt, E. S., Acquarini, E., y Di Ceglie, A. (2002). Psychometric properties of an adolescent version of the barratt impulsiveness scale-11 for a sample of italian high school students. *Perceptual and Motor Skills, 95*(2), 621-635.

Fossati, A., Ceglie, A. D., Acquarini, E., y Barratt, E. S. (2001). Psychometric properties of an italian version of the barratt impulsiveness scale-11 (BIS-11) in nonclinical subjects. *Journal of Clinical Psychology, 57*(6), 815-828.

French, B. F., y Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling, 13*(3), 378-402.

French, A. W. y Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33* (3), 315-333.

Gómez, J. (1996). Aportaciones de los modelos de estructuras de covarianza al análisis psicométrico. En J. Muñiz (Ed.), *Psicometría* (pp. 457-554). Madrid: Universitas, S.A.

Gómez, J. e Hidalgo, M. D. (1997). Evaluación del funcionamiento diferencial en ítems dicotómicos: Una revisión metodológica. *Anuario de Psicología, 74*, 3-32.

Gómez, J., Hidalgo, M. D., y Guilera, G. (2010). El sesgo de los instrumentos de medición. Tests justos. *Papeles del Psicólogo, 31*(1), 75-84.

Gómez, J., y Navas, M. J. (1998). Impacto y funcionamiento diferencial de los ítems respecto al género en una prueba de aptitud numérica. *Psicothema, 10*(3), 685-696.

Green, S. B., y Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika, 74*(1), 121-135.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*, 427-438.

Gulliksen, H. y Wilks, S. S. (1950). Regression tests for several samples. *Psychometrika*, 15, 91-114.

Haberman, J.S. (1977). Log-linear models and frequency tables with small expected cell

counts. Annals of Statistics, 5, 1148-1169.

Hair, J. F., Anderson, R. E., Tatham, R. L., y Black, W. C. (1999). *Análisis multivariante* (5ª ed.). Madrid: Prentice Hall Ibérica.

Haladyna, T. M., y Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. En R. L. Linn (Ed.), *Educational measurement (3rd ed.)*. (pp. 147-200). New York: Macmillan Publishing Co, Inc; American Council on Education.

Hambleton, R. y Swaminathan, H. (1985) Item Response Theory. Principles and applications. Boston: Kluwer Nijhoff Publishing

Hambleton, R. K., Swaminathan, H. y Rogers, H. J. (1991). *Fundamentals of item response theory*. Thousand Oaks, CA, US: Sage Publications, Inc.

Harmon-Jones, E.; Barratt, E. S. y Wigg, C. (1997). Impulsiveness, aggression, reading, and the p300 of the event-related potential. *Personality and Individual Differences*, 22 (4), 439-445.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11*(2), 91-115.

Hart, S. D., y Dempster, R. J. (1997). Impulsivity and Psychopathy. En C. D. Webster y M. A. Jackson (Eds.), *Impulsivity; theory, assesment and treatment* (pp. 212–232). New York: Guilford Press.

Harvey, R., y Hammer, A. (1999). Item response theory. *The Counseling Psychologist*, *27*(3), 353-383.

Hidalgo, M. D. y Gómez, J. (1999). Técnicas de detección del funcionamiento diferencial en ítems politómicos. *Metodología de las Ciencias del Comportamiento, 1*, 39-60.

Hidalgo, M. D. y López, J. A. (2000). Funcionamiento diferencial de los ítems: Presente y perspectivas de futuro. *Metodología de las Ciencias del Comportamiento, 2*, 167-182.

Holgado, F. P., Chacón, S., Barbero, I., y Vila, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity: International Journal of Methodology, 44*(1), 153-166.

Holland, P. W. y Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel Procedure. En H. Wainer y H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, N. J.: Lawrence Erlbaum Associates.

Holland, P. W. y Wainer, H. (Eds.) (1993). *Differential Item Functioning*. Hillsdale, NJ: LEA.

Horn, J. L. (1991). Comments on "issues in factorial invariance". En L. M. Collins, y J. L. Horn (Eds.), *Best methods for the analysis of change* (pp. 114-125). Washington DC: American Psychological Association.

Horn, J. L. y McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18* (*3-4*), 117-144.

Horn, J. L., McArdle, J. J., y Mason, R. (1983). When is invariance not invarient: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist, 1*(4), 179-188.

Hoyle, R. H. e. (1995). *Structural equation modeling: Concepts, issues and applications*. Thousand Oaks, CA, US: Sage Publications, Inc.

Hu, L. T. y Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1-55.

Hui, C. H., y Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology, 16*(2), 131-52.

Jensen, A. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, *39*(1), 1-123.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Johnson, D. R., y Creech, J. C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review, 48*(3), 398-407.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109-133.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. En A. S. Goldberger y O. D. Duncan (Eds.) *Structural equation models in the Social Sciences*. Nueva York: Seminar.

Jöreskog, K. G. (1990). New development in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity, 24*, 387-404.

Joreskog, K. G. (1993). Testing structural equation models. En K. A. Bollen y J. S. Long, *Testing structural equation models* (pp. 294-316). Newbury Park, CA: Sage.

Joreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika, 59*(3), 381-89.

Jöreskog, K. G. (2002). Structural equation modeling with ordinal variables using LISREL. Extraído en Julio de 2005 de http://www.ssicentral.com/lisrel/techdocs/ordinal.pdf

Jöreskog, K. G., y Sörbom, D. (1979). *Advances in factorial analysis and structural equation models*. Cambridge, MA: Abt Books.

Jöreskog, K. G. y Sörbom, D. (1989). *LISREL 7 User's reference guide*. Chicago: Scientific Software, Inc.

Jöreskog, K. G. y Sörbom, D. (1996). *LISREL 7: User´s reference guide*. Chicago: Scientific Software.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educatinal measurement* (4ª ed., pp. 17-64). Wesport, CT: American Council on Education and Praeger Publishers.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. California: Sage Publications, Inc.

Kim, S., y Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education, 8*(4), 291-312.

Kim, S. H., y Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22*(4), 344-355.

Kim, S.H., Cohen, A.S., y Park, T.H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement, 32*, 261-276.

Kim, S., Kim, S. H., y Kamphaus, R. (2010). Is aggression the same for boys and girls? Assessing measurement invariance with confirmatory factor analysis and item response theory. *School Psychology Quarterly*, *25*(1), 45-61.

Kim, E. S., y Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling, 18*(2), 212-228.

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*(5), 746-759.

Kline, P. (1998). *The new psychometrics: Science, psychology and measurement.* Londres: Routledge.

Knorring, L., y Ekselius, L. (1998). Psychopharmacological treatment and impulsivity. En T. Millon, E. Simonsen, M. Birket-Smith, y R. D. Davis (Eds.), *Psychopathy, antisocial, criminal and violent behaviour*. London: Guilford Press.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*(11), 1183-1192.

Labouvie, E. y Ruetsch, C. (1995). Testing for equivalence of measurement scales: Simple structure and metric invariance reconsidered. *Multivariate Behavioral Research, 30* (*1*), 63-76.

Lautenschlager, G. L. y Park, D. G. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness, and parameter linking. *Applied Psychological Measurement, 12* (4), 365-376.

Lawley, D. N. (1943). The application of the maximum likelihood method to factor analysis. *British Journal of Psychology, 33*, 172-175.

Lévy J. P., Varela J. (Eds.) (2006). *Modelización con estructuras de covarianzas en ciencias sociales. Temas esenciales, avanzados y aportaciones especiales*. A Coruña: Netbiblo.

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. En P. W. Holland, y H. Wainer (Eds.), *Differential item functioning* (pp. 349-364). New Jersey: Lawrence Erlbaum Associates.

Little, T.D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*, 53-76.

Little, T. D. (2000). On the comparability of constructs in cross-cultural research: A critique of Cheung and Rensvold. *Journal of Cross-Cultural Psychology, 31* 213-219.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635-694.

López-Pina, J. A., e Hidalgo, M. D. (1996). Bondad de ajuste y teoría de respuesta a los ítems. En J. Muñiz (Ed.), *Psicometría* (pp. 643-703). Madrid: Universitas.

Lord, F. M. (1952). *A theory of tests scores.* Iowa City, IA: Psychometric Society.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. y Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lozano, L. M., García-Cueto, E., y Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 4*(2), 73-79.

Lubke, G. H., y Dolan, C. V. (2003). Can unequal residual variances across groups mask differences in residual means in the common factor model? *Structural Equation Modeling, 10*(2), 175-192.

MacCallum, R. C. (1995). Model especification: Procedures, strategies, and related issues. En R. H. Hoyle, *Structural equation modeling: Concepts, issues, and applications* (pp. 76-99). Newbury Park, CA: Sage.

MacCallum, R. C., Roznowski, M., Mar, M. y Reith, J. V. (1994). Alternative strategies for cross-validation of covariance structure models. *Multivariate Behavioral Research, 29*, 1-32.

MacCallum, R. C., Roznowski, M. y Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*, 490-504.

Manzano, V. (1998). La calidad del muestreo en las investigaciones sociales. *Revista Electrónica de Metodología Aplicada, 3*(1), 16-29.

Marsh, H. W., Balla, J. R., y McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*(3), 391-410.

Marsh, H. W., Hau, K. T., Balla, J. R. y Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33* (2), 181-220.

Martínez, M. R., Hernández, M. J., y Hernández, M. V. (2006). *Psicometría*. Madrid: Alianza Editorial.

Maurer, T. J., Raju, N. S., y Collins, W. C. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology,* 83, 693-702.

Maydeu-Olivares, A., Drasgow, F., y Mead, A. D. (1994). Distinguishing among parametric item response models for polychotomous ordered data. *Applied Psychological Measurement, 18* (*3*), 245-256.

Maydeu-Olivares, A., Morera, O. y D´Zurilla, T. J. (1998). Using graphical methods in assessing measuremet invariance in inventory data. *Multivariate Behavioral Research*, *34*(3), 397-420.

McArdle, J. J. y McDonald, R. P. (1984). Some algebraic properties of the reticular model for moment structures. *British Journal of Mathematical and Statistical Psychology, 37* (2), 234-251.

McDonald, R. P. (1999). *Test theory: A unified treatment*. New Jersey: Lawrence Erlbaum Associates, Publishers.

McDonald, R. P. y Mok, M. M. (1995). Goodness of fit in item response models. Multivariate Behavioral Research, 30, 23-40.

McIntire, S., y Miller, L. (2007). *Foundations of psychological testing: A practical approach (2nd ed.)*. Thousand Oaks, CA US: Sage Publications, Inc.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95*(4), 728-743.

Meade, A. W., Johnson, E. C., y Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568-592.

Meade, A. W. y Lautenschlager, G. J. (2004a). A Comparison of Item Response Theory and Confirmatory Factor Analytic Methodologies for Establishing Measurement Equivalence/lnvariance. *Organizational Research Methods 7* (4), 361-388.

Meade, A. W. y Lautenschlager, G. J. (2004b). A monte-carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling, 11* (1), 60-72.

Meade, A. W. y Lautenschlager, G. J. (2004c). *Same question, different answers: CFA and two IRT approaches to measurement invariance*. Symposium presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chichago, IL.

Meade, A., Lautenschlager, G., y Johnson, E. (2007). A Monte Carlo examination of the sensitivity of the differential functioning of items and tests framework for tests of measurement invariance with Likert data. *Applied Psychological Measurement*, *31*(5), 430-455.

Mellenberg, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105-108.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19*(1), 91-100.

Meredith, W. (1964). Notes of factorial invariance. *Psychometrika,* 177-185.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance.

*Psychometrika, 58*, 525-543.

Meredith, W. (1995). Two wrongs still do not make a right. *Multivariate Behavioral Research, 30* (*1*), 117.

Meredith, W. y Horn, J. (2001). The role of factorial invariance in modeling growth and change. En L. M. Collins y A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 203-240). Washington DC: American Psychological Association.

Meredith, W. y Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 58* (2), 289-311.

Meredith, W., y Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44*(11 Suppl 3), S69-77.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*(10), 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012-27.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. I. Braun (Ed.), *Test validity.* (pp. 33-48). Hillsdale, NJ England: Lawrence Erlbaum Associates, Inc.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (3rd ed.).* (pp. 13-103). New York, NY England: Macmillan Publishing Co, Inc; American Council on Education.

Milfont, T. L., Duckitt, J., y Cameron, L. D. (2006). A cross-cultural study of environmental motive concerns and their implications for proenvironmental

behavior. *Environment and Behavior, 38*(6), 745-767.

Milfont, T. L., y Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3*(1), 111-121.

Miller, M. D. y Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement, 16* (4), 381-388.

Miller, T., y Spray, J. (1993). Logistic Discriminant Function Analysis for DIF Identification of Polytomously Scored Items. *Journal of Educational Measurement*, *30*(2), 107-22.

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research, 30* (*4*), 577-605.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychologicals Methods, 2* (3), 248-260.

Millsap, R. E. (1998). Group differences in regression intercepts: Implications for factorial invariance. *Multivariate Behavioral Research, 33*(3), 403-424.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY US: Routledge/Taylor & Francis Group.

Millsap, R. E. y Everson, H. T. (1993). Metodology review: Statistical approaches for measuring test bias. *Applied Psychological Measurement*, *17*(4), 297-334.

Millsap, R. E. y Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9*(1), 93-115.

Millsap, R. E., y Meredith, W. (2007). Factorial invariance: Historical perspective and new

problems. En R. Cudeck y R. C. MacCallum (Eds.), Factor analysis at 100: Historical developments and future directions (pp. 131–152). Mahwah, NJ: Erlbaum.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11*, 3-31.

Moeller, F.G., Dougherty, D.M., Barratt, E.S., Schmitz, J.M., Swann, A.C. y Grabowski, J. (2001). The impact of impulsivity on cocaine use and retention in treatment. *Journal of Substance Abuse and Treatment, 21*, 193-198.

Mulaik, S. A. (1986). Factor analysis and psychometrika: Major developments. *Psychometrika, 51*(1), 23-33.

Muñiz, J. (2001). Estatus métrico de las puntuaciones. En J. Muñiz, *Teoría clásica de los tests* (pp. 281-302). Madrid: Pirámide.

Muñiz, J. (2004). La validación de los tests. *Metodología De Las Ciencias Del Comportamiento, 5*(2), 121-141.

Muraki, E. (1996). A generalized partial credit model. En W. J. van der Linden y R. K. Hambleton, *Handbook of modern item response theory*. New York: Springer-Verlag.

Muthen, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*, 115-132.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29* (*1*), 81-117.

Muthén, B., y Kaplan, D. (1985). A comparison of some methodologies for the factor

analysis of nonnormal likert variables. *British Journal of Mathematical and Statistical Psychology, 38*, 171-189.

Navas, M. J. (1997). *Proyecto docente de Psicometría*. Madrid: UNED.

Navas, M. J. (2001). Métodos, Diseños y Técnicas de Investigación en Psicología. Madrid: UNED.

Olsson, U. (1979). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research, 14*, 485-500.

Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika, 6* (*2*), 150-166.

Oquendo, M.A., Baca-García, E., Graver, R., Morales, M., Montalvan, V., and Mann, J.J. (2001). Spanish Adaptation of the Barratt Impulsiveness Scale (BIS-11). *European Journal of Psychiatry, 15* (*3*), 147-155.

Oshima, T. C., Kushubar, S., Scott, J.C. y Raju N.S. (2009). *DFIT8 for Window User's Manual: Differential functioning of items and tests.* St. Paul MN: Assessment Systems Corporation.

Oshima, T. C., y Morris, S. B. (2008). An NCME instructional module on Raju's Differential Functioning of Items and Tests (DFIT). *Educational Measurement: Issues and Practice, 27*(3), 43-50.

Oshima, T. C., Raju, N. S., y Flowers, C. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests. *Journal of Educational Measurement, 34*, 253-272.

Oshima, T. C., Raju, N. S., y Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT)

framework. *Journal of Educational Measurement, 43*(1), 1-17.

Patton, J.H., Stanford, M.S., y Barratt, E.S. (1995). Factor Structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psychology, 51* (*6*), 768-774.

Penfield, R. D. (2005). DIFAS: Differential item functioning analysis system. *Applied Psychological Measurement, 29*(2), 150-151.

Penfield, R. D. (2010). Distinguishing between net and global DIF in polytomous items. *Journal of Educational Measurement, 47*(2), 129-149.

Penfield, R. D., y Camilli, G. (2007). Differential item functioning and item bias. En C. R. Rao, y S. Sinharay (Eds.), *Handbook of statistics vol. 26* (, pp. 125-167). Amsterdam: Elsevier.

Penfield, R. D., y Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*(3), 5.

Potenza, M. T. y Dorans, N. J. (1995). DIF Assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23-37.

Prieto, G., y Delgado, A. R. (2010). Fiabilidad y validez. *Papeles Del Psicólogo, 31*(1), 67-74.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*,

197-207.

Raju, N. S. y Ellis, B. B. (2002). Differential item and test functioning. En F. Drasgow y N. Schmitt, *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis.* San Francisco, C. A.: Jossey-Bass.

Raju, N. S., Fortmann-Johnson, K., Kim, W., Morris, S. B., Nering, M. L., y Oshima, T.C. (2009). The item parameter replication method for detecting differential functioning in the polytomous DFIT framework. *Applied Psychological Measurement, 33*(2), 133-147.

Raju, N. S., Laffitte, L. J., y Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87* (3), 517-529.

Raju, N. S., van der Linden, W., y Fleer, P. (1995). An IRT-based internal measure of test bias with applications for differential item functioning. *Applied Psychological Measurement, 19*, 353-368.

Rasch (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute for Educational Research.

Recio, P., Santisteban, C. y Alvarado, J.M. (2004). Estructura factorial de una adaptación española del test de impulsividad de Barrat. *Metodología de las Ciencias del Comportamiento, Suplemento,* 515-519.

Reckase, M. D. (1979). Unifactor latent trait models applied to multi-factor tests: results and implications. *Journal of Educational Statistics, 4*, 207-230.

Reise, S. P., Widaman, K. F., y Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance.

*Psychological Bulletin, 114*, 552-566.

Revuelta, J., Abad, F. J., y Ponsoda, V. (2006). *Introducción a los modelos politómicos de la teoría de respuesta al ítem*. Madrid: La Muralla

Reynolds, C. R. y Brown, R. T. (Eds.) (1984). *Perspectives on bias in mental testing*. New York: Plenum Press.

Riordan, C. M., Richardson, H. A., Schaffer, B. S., y Vandenberg, R. J. (2001). Alpha, beta and gamma change: A review of past research with recommendations for new directions. En C. A. Schriesheim y L. L. Neider (Eds). *Equivalence of measurement.* Greenwich, CT: Information Age Publishing.

Riordan, C. M. y Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management, 20*, 643-671.

Robie, C., Zickar, M. J., y Schmit, M. J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance, 14*(2), 187-207.

Rock, D. A., Werts, C. E., y Flaugher, R. L. ( 1978). The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations. *Multivariate Behavioral Research*, *13*, 403-418.

Roznowski, M. (1987). Use of tests manifesting sex differences as measures of intelligence: Implications for measurement bias. *Journal of Applied Psychology, 72*, 480-483.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*, 100-114.

Samejima, F. (1997). Graded response model. En W. J. van der Linden y R. K. Hambleton, *Handbook of modern item response theory*. New York: Springer-Verlag.

Santisteban, C. (1990). *Psicometría: Teoría y práctica en la construcción de tests*. Madrid: Ediciones Norma.

Santisteban, C. (2009): *Principios de Psicometría*. Madrid: Síntesis.

Santisteban, C. y Alvarado, J. M. (2001). *Modelos psicométricos*. Madrid: UNED.

Santisteban, C., Alvarado, J. M. y Recio, P. (2007). Evaluation of a spanish version of the Buss and Perry aggression questionnaire: Some personal and situational factors related to the aggression scores of young subjects. *Personality and Individual Differences, 42* (8), 1453-1465.

Savage, L. W., y Ehrlich, P. (1990). *Philosophical and foundational issues in measurement theory*. Hillsdale, N. J. Lawrence Erlbaum Associates.

Schaubroeck, J. y Green, S. G. (1989). Confirmatory factor analytic procedures for assessing change during organizational entry. *Journal of Applied Psychology*, *74*, 892-900.

Schmitt, N. (1982). The use of covariance structures to assess beta and gamma change. *Multivariate Behavioral Research 1*7, 343-358.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350-353.

Schmit, M. J., Kihm, J. A., y Robie, C. (2000). Developmet of a global measure of personality. *Personnel Psychology 53*(1), 153-193.

Shealy, R., y Stout, W. (1993). An item response theory model for test bias and differential test functioning. *Differential item functioning* (pp. 197-239). Hillsdale, NJ England: Lawrence Erlbaum Associates, Inc.

Sijtsma, K. (2009a). Correcting fallacies in validity, reliability, and classification. *International Journal of Testing, 9*(3), 167-194.

Sijtsma, K. (2009b). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107-120.

Smith, L. L. y Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction scale. *Journal of Personality and Social Psychology, 75* (*5*), 1350-1362.

Soloff, P.H., Kelly, T.M., Strotmeyer, S.J., Malone, K.M. y Mann, J.J. (2003). Impulsivity, gender and response to fenfluramine challenge in borderline personality disorder. *Psychiatry Research, 119*, 11-24.

Someya, T., Sakado, K., Seki, T., Kojima, M., Reist, C., Tang, S. W., y Takahashi, S. (2001). The japanese version of the barratt impulsiveness scale, 11th version (BIS-11): Its reliability and validity. *Psychiatry and Clinical Neurosciences, 55*(2), 111-114.

Stark, S. (2001). MODFIT [Computer Software]. Descargado en Octubre de 2004 de http://io.psych.uiuc.edu/irt.

Stark, S., Chernyshenko, O. S., y Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292-1306.

Stark, S., Chernyshenko, O. L., Lancaster, A. R. , Drasgow, F., y Fitzgerald, L. F. (2002). Toward standardized measurement of sexual harassment: Shortening the SEQ-DoD using item response theory. *Military Psychology, 14* (*1*), 49-72.

Steenkamp, J. E. M. y Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research 25*, 78-90.

Steiger, J. H. (1990). Structural model evaluation an modification: An interval estimation approach. *Multivariate Behavioral Research, 25* (2), 173-180.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. En S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1-49). New York: Wiley.

Spearman, C. (1904). 'General intelligence,' objectively determined and measured. *The American Journal of Psychology, 15*(2), 201-293.

Swaminathan, H., Hambleton, R. K., y Rogers, H. J. (2007). Assessing the fit of item response theory models. En C. R. Rao, y S. Sinharay (Eds.), *Handbook of statistics vol. 26*. Amsterdam: Elsevier.

Swaminathan, H. y Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Swann, A.C., Pazzaglia, P., Nicholls, A., Dougherty, D.M. and Moeller, F.G. (2003). Impulsivity and phase of illness in bipolar disorder. *Journal of Affective Disorders, 73*, 105-111.

Tanaka, J. S. (1987). "How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. *Child Development, 58* (1), 134-146.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. En K. A. Bollen, y J. S. Long (Eds.), *Testing structural equation models.* (pp. 10-39). Thousand Oaks, CA: Sage Publications, Inc.

Tanaka, J. S. y Huba, G. J. (1984). Confirmatory hierarchical factor analysis of psychological distress measures. *Journal of Personality and Social Psychology, 46*, 621-635.

Taris, T. W., Bok, I. A., y Meijer, Z. Y. (1998). Assessing stability and change of psychometric properties of multi-item concepts across different situations: A general approach. *Journal of Psychology: Interdisciplinary & Applied 132*(3), 301-316.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Cook, K. F., Crane, P. K., Gibbons, L. E.,... Cella, D. (2007). Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): Applications (with illustrations) to measures of physical functioning ability and general distress. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 16*(1), 43-68.

Thissen, D. (2001). *IRTLRDIF v2.02b: Software for the computation of the statistics involved in item response theory likelihood-ratio test for differential item functioning (Computer software)*. Chapel Hill, NC: LL Thurstone Psychometric Laboratory.

Thissen, D. (1991). *MULTILOG users guide: Multiple categorical item analysis and test scoring using item respone theory (Computer software)*. Chicago: Scientific Software International.

Thissen, D. y Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567-577.

Thissen, D., Steinberg, L. y Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99* (*1*), 118-128.

Thissen, D., Steinberg, L., y Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. En H. Wainer y H. I. Braun. *Test validity*. 147-169. Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., y Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. En P. W. Holland y H. Wainer (Eds.) *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Thurstone, L. L. (1947). *Multiple factor analysis.* University of Chicago Press: Chicago.

Tinsley, H. E. y Tinsley, D. J. (1987). Uses of factor analysis in counseling psychology research. *Journal of Counseling Psychology 34*, 414-424.

Tomás, I., González-Romá, V., y Gómez, J. (2000). Teoría de respuesta al ítem y análisis factorial confirmatorio: Dos métodos para analizar la equivalencia psicométrica en la traducción de cuestionarios. *Psicothema, 12*(2), 540-544.

Tomás, J. M., y Oliver, A. (2004). Análisis Psicométrico Confirmatorio de una Medida Multidimensional del Autoconcepto en Español. *Revista Interamericana de Psicología, 38*(2), 285-293.

Tremblay, R. E., Pihl, R. O., Vitaro, F., y Dobkin, P. L. (1994). Predicting early onset of male antisocial behavior from preschool behavior. *Archives of General Psychiatry, 51*, 732–739.

Van de Vijver, F. J. y Poortinga, Y. H. (1982). Cross-cultural generalization and universality. *Journal of Cross-Cultural Psychology, 13*(4), 387-408.

Vandenberg, R. J. (2002). Toward a further understanding of an improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5* (2), 139-158.

Vandenberg, R.J. y Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3* (1), 4-69.

Van der Linden, W., y Hambleton, R. K. (1997). *Handbook of modern item respnse theory*. Nueva York: Springer

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. En P. W. Holland y H. Wainer (Eds). *Differential item functioning*. (pp.123-135). Hillsdale, NJ: Erlbaum.

Welch, C., y Hoover, H. (1993). Procedures for Extending Item Bias Detection Techniques to Polytomously Scored Items. *Applied Measurement in Education*, *6*(1), 1-19.

Weng, L. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64*(6), 956-972.

West, S. G., Finch, J. F., y Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Procedures, concepts, issues and applications.* (pp. 56-75). Newbury Park: Sage.

Widaman, K. F. y Reise, S. P. (1997). Exploring the measurement invariance of

psychological instruments: Applications in the substance use domain. En K. J. Bryant, M. Windle, y S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281-324). Washington DC: American Psychological Association.

Wilmut, J. (1975). Objective test analysis: some criteria for item selection. *Research in Education, 13*, 27-56.

Wu, A.D., Li, Z., y Zumbo, B.D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation, 12*(3), 1-26.

Zickar, M. J. y Robie, C. (1999). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology, 84* (4), 551-563.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. En P. W. Holland, y H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). New Jersey: Lawrence Erlbaum Associates.

Zieky, M. (2006). Fairness reviews in assessment. En T. M. Haladyna (Ed.), *Handbook of test development.* (pp. 359-376). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? implications for translating language tests. *Language Testing, 20*(2), 136-147.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. En C. R. Rao, y S. Sinharay (Eds.), *Handbook of statistics vol. 26. psychometrics* (pp. 45-79). Amsterdam: Elsevier.

Zumbo, B. D., Gadermann, A. M., y Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for likert rating scales. *Journal of Modern Applied Statistical Methods, 6*, 21-29.

Zumbo, B. D., y Rupp, A. A. (2004). Responsible modelling of measurement data for appropriate inferences: Important advances in reliability and validity theory. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 73-92). Thousand Oaks, CA: Sage Press.

Zwick, R. y Thayer, D. T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics, 21* (3), 187-201.

Zwick, R., Thayer, D., y Mazzeo, J. (1997). Descriptive and Inferential Procedures for Assessing Differential Item Functioning in Polytomous Items. *Applied Measurement in Education*, *10*(4), 321-344.

# Anexos

**Anexo 1.** Ítems de la Escala de Impulsividad de Barratt Adaptada (BIS)

## ESCALA DE IMPULSIVIDAD DE BARRAT ADAPTADA

Las frases que aparecen a continuación se refieren a diferentes formas de actuar y de pensar. Lee atentamente cada una de ellas y pon un aspa (X) en la respuesta que más se ajusta a tu forma de ser. MUCHAS GRACIAS.

| | Nunca o casi nunca | Algunas veces | Bastantes veces | Siempre o casi siempre |
|---|---|---|---|---|
| 1. Cuando voy a hacer algo, lo preparo muy bien antes. | | | | |
| 2. Hago cosas sin pensar. | | | | |
| 3. Soy despreocupado/a, distraído/a. | | | | |
| 4. Pienso muchas cosas a la vez. | | | | |
| 5. Hago mis planes con mucho tiempo. | | | | |
| 6. Aunque me digan que espere para abrir un regalo no hago caso. | | | | |
| 7. Me concentro fácilmente. | | | | |
| 8. Ahorro regularmente. | | | | |
| 9. Me resulta difícil permanecer sentado/a o callado/a durante mucho tiempo. | | | | |
| 10. Soy un/a chico/a que piensa bastante las cosas. | | | | |
| 11. Me preocupo por obtener buenas notas. | | | | |
| 12. Digo cosas sin pensar. | | | | |
| 13. Me gusta pensar en cosas que me parece que son difíciles. | | | | |
| 14. Me canso de los deberes de una asignatura y empiezo los de otra sin haber terminado los primeros. | | | | |
| 15. Me dicen que hago las cosas de manera un poco alocada. | | | | |
| 16. Me aburro fácilmente cuando tengo que resolver problemas que exigen pensar mucho. | | | | |
| 17. Me preocupa estar enfermo. | | | | |
| 18. Hago las cosas de pronto, sin pensar. | | | | |
| 19. Me pienso bastante todo. | | | | |
| 20. Me canso enseguida de todo. | | | | |
| 21. Compro cosas dejándome llevar por mis impulsos. | | | | |
| 22. Acabo lo que empiezo. | | | | |
| 23. Me muevo y ando más rápido que mis amigos. | | | | |
| 24. Resuelvo los problemas como primero se me ocurre. | | | | |
| 25. Intento comprar cosas más caras del dinero que tengo. | | | | |
| 26. Hablo rápido en comparación con mis amigos. | | | | |
| 27. Pienso en cosas raras. | | | | |
| 28. Estoy más interesado en el presente que en el futuro. | | | | |
| 29. Estoy nervioso en clase. | | | | |

| | | | | |
|---|---|---|---|---|
| 30. Hago planes para cuando sea mayor. | | | | |

**Anexo 2.** Instrucciones para los encuestadores

INSTRUCCIONES EN LA APLICACIÓN DE LOS TESTS

[*Primera sesión*:]

Buenos días, mi nombre es “ ” y el de mi compañero “ ”.

Participamos en un estudio financiado por el Ministerio de Educación y Cultura, para el que recogemos datos en diversos colegios.

A continuación os voy a entregar un cuadernillo. Lo primero que tenéis que hacer es escribir en la portada el número que os ha entregado mi compañero.

¿Ya lo habéis apuntado todos?

Si alguno de vosotros no habla bien el castellano tiene que ponerlo en el cuadernillo, debajo del número que haya apuntado, escribiendo también el idioma que se hable en su casa.

Bien, ahora abrid el cuadernillo; como veis, consta de 5 hojas [*7 en el caso de alumnos de 15-16 años*].

En esas hojas hay muchas preguntas, a las que debéis contestar con sinceridad. De que seáis sinceros depende que todo este trabajo que se hace sirva para que se puedan mejorar las relaciones entre las personas.

No tengáis ningún problema en ser realmente sinceros, porque como los cuestionarios son anónimos, no llevan nombre y nadie sabrá a quién pertenece cada respuesta.

Hay varias partes, y antes de comenzar cada parte os leeremos las instrucciones para deciros la forma de contestar y os pondremos un ejemplo.

Vamos a empezar:

[*Se leen literalmente las instrucciones del primer cuestionario. Una vez leídas hay que poner un ejemplo sobre la forma de responder y preguntarles si lo han entendido bien*]. [*Esta operación se repite con todos los cuestionarios*].

En el caso de que no entendáis alguna pregunta dejadla en blanco. Ya podéis empezar.

[*Segunda sesión:*]

El cuestionario que os ha entregado mi compañero consta de 4 folios más la portada. Son en total, 39 preguntas, a las que tenéis que contestar con sinceridad; también es anónimo, por lo que no olvidéis apuntar vuestro número en la portada del cuadernillo. Ya podéis empezar…

[*Hay que asegurase de que apuntan el número en los cuadernillos*]